

5 Person, Personal Identity, and Personality

5.1 Preliminaries

It was mentioned in the introduction that one of the major types of concerns related to new methods of intervening in the brain is expressed in the question whether the identity of those persons they are applied to is put at risk. Worries of this kind have already been voiced in the past with respect to traditional intervention techniques used in neurosurgery, psychopharmacology, and even (non-invasive) psychotherapy. However, the availability of new techniques like neural grafting, neuroprosthetics and electrical brain stimulation endows those old concerns with new socio-political vigour.²² In this preliminary section we will attempt to disambiguate the fears of those who would regard certain techniques for intervening in the brain as a potential threat to personal identity. In doing so, it will turn out that, while some particular worries falling within this category need not be dealt with at great length, others require a clarification of the concepts of identity in general, of personal identity in particular, and of related concepts like personhood and personality before a comprehensible assessment of possible hazards can be given for different techniques.

As an example of the type of concern that can be dismissed relatively easily, we bring to mind a rather peculiar worry pertaining specifically to the technique of neural grafting: Some people seem to worry that certain of a donor's personality traits, which might be encoded in a neural graft, could get transferred to the graft recipient. The reason why this particular concern can be put aside without much ado is that it is based on an all too naïve picture of the brain's functioning, on the one hand, and of the technique of neural grafting, on the other. Even though, in a sense, there is truth to the claim that certain delimited parts of the human brain are specifically involved in the implementation of certain psychic functions, the brain doesn't work like a construction kit with neatly separable modules for each of a person's functional traits. In any case, such brain parts could certainly not confer "alien" personality traits when dissected from *immature* brains and inserted in a topologically analogous *mature*, but malfunctioning site in the skull of a person (Boer 1999:470). But *this* (not a transplantation of brain parts of other persons) is exactly what is done through the technique of neural grafting. Considering, for instance, the actual experimental treatment of

²² While this concern pertains *mutatis mutandis* to all new techniques for intervening in the psyche, in recent years it has been widely discussed especially with regard to brain tissue transplantation. See Boer 1994, 1996, 1999; Franz 1996; Hildt 1996, 1999a, 1999b; Jelden 1996; Linke 1993; Metzinger 1996; Northoff 1995, 1996a, 1996b, 2001; Quante 1996.

Parkinson's disease by means of neural grafts, the danger of alien thoughts, memories, preferences, etc. intruding a recipient's mind and undermining his identity becomes utterly elusive since the sole source for acquiring neuronal grafts at present are electively aborted human foetuses or pig foetuses.

However, even if there is no reason to expect a donor's psychic traits encroaching upon the personality of the respective recipient, the inserted material might still act as a disintegrating force on a recipient's personal identity – in the sense of leading either to a breakdown of personhood itself or to an alteration of his personality so radical that, in the end, he would be “no longer the same”. Let's take a closer look at both of these possible perils²³ in turn. The expression “breakdown of personhood” is meant to describe a situation in which following an intervention a patient can no longer be said to be a person. Whether one takes this to be an actual possibility will depend on one's general understanding of personhood. For there is an old tradition according to which the use of the concept of a person doesn't allow for human beings to be deprived of personhood. This tradition, which is deeply rooted in medieval theological thinking, acknowledges the possibility of nonhumans being persons (e.g. angels), but on the other hand holds that there can be no human being which for any reason or at any time in its life is not to be considered as a person.²⁴ According to a second – distinctively modern and secular – strand of tradition, though, human beings do not qualify in and of themselves as persons, but rather by virtue of certain capacities. We will see later on that there is some disagreement among the proponents of this view as to *which* capacities are those that are essential for personhood. However, usually these capacities are taken to be such that no human being already possesses them at the time of birth (s. e.g. Quante 2002:20) and that for any person there is some likelihood to lose these capacities again (either permanently or for some period during their lives).

To illustrate the dissent between the two traditions, let us consider, for a moment, the case of human beings living in a *persistent vegetative state* (PVS). Being in such a wakeful yet unconscious state, they more or less lack all the characteristics (like e.g. awareness and self-consciousness) deemed to be constitutive for being a person by followers of the second tradition. Still, adherents of the first tradition would take offence at the claim that humans trapped in PVS are not to be considered as persons. The reason for their uneasiness is that they conceive of personhood as the distinctive status of beings deserving moral concern. Given this assumption, conceding that human beings lose personhood by entering into PVS implies that they forfeit

²³ Please note that “possible” here does just mean that the occurrence of such effects cannot be ruled out by a priori arguments. It does *not* mean that they are empirically established “possible side effects” of the technique.

²⁴ Accordingly, most contemporary philosophers who subscribe to this view have a strong affinity to theology, see for instance Ford (2002) and Spaemann (1996).

their right to respectful treatment.²⁵ The idea of (living) human beings devoid of any rights and discriminated against in medical care is as unacceptable to us as it is to the objectors of non-personal human beings. Still, in this chapter we will develop an account of personhood along the lines of the second tradition, that is, we acknowledge the possibility of human beings which are not persons. The reason for this is that for a *serious* consideration of concerns regarding the identity of persons that undergo a new kind of medical intervention this approach is without alternative. After all, if one presupposes along the lines of the first tradition mentioned above that “person” is just another word for “human being”, then it becomes a trivial truth that nothing that you can do with a person’s psyche will change her identity. For no matter how thoroughly you change the personality of a human being, it will still remain the same individual of the species *homo sapiens sapiens*. However, to forestall misconceptions we would like to emphasise from the outset that we do *not* require holders of moral or legal rights to be persons. Instead, we will argue that personhood is a prerequisite just for the ascription of *duties*, but not for the recognition of *rights*.

Coming back to where we started, the expression “breakdown of personhood” can be used to refer to the situation of patients that – due to some medical intervention or for other reasons – end up being no longer a person. Of course, if there was a higher than negligible risk for patients undergoing a particular kind of intervention to be bereft of necessary capabilities for being regarded as persons, then the procedure in question would not be employed. In this respect, science and medical practice has obviously learned from the history of psychosurgical interventions in the brain. For instance, in the early 1950s lobotomies were still performed at a rate of 5.000 per year in the United States notwithstanding their side effects, which include inertia, apathy, decreased attention, social inappropriateness, and seizures (Missa 1998:737). How drastic a change in personality can result from brain surgery has been famously depicted by Jack Nicholson in Milos Forman’s movie “One Flew Over the Cuckoo’s Nest” (1975). Nicholson played McMurphy, a rebellious patient in a psychiatric ward, who in the end is subdued by lobotomy, thereby turned into an apathetic wreck.

With Nicholson’s performance in mind one might wonder if a person’s identity can get “extinguished” without it being replaced by a new one, but also without the person ceasing to exist altogether as is the case with PVS. Let us suppose that after the lobotomy McMurphy still knows who and where he is though he doesn’t care much about it, furthermore, that he is still able to

²⁵ This would then, of course, include medical care. On this background, especially considering the crimes committed by physicians under the Third Reich, the concern of medical practitioners is fully understandable: Normatively loaded terms like “person” could potentially be misused to discriminate against certain patients, and should therefore, in practice, be either avoided or applied indiscriminately to all human beings (as potential patients).

talk and reason though usually he doesn't feel like doing so. According to some theories of personhood he thereby would still fulfil the minimal requirements for being a person. What he lacks after the operation, so one could argue, are all the distinguishing traits that defined his former identity. Moreover, as he didn't acquire any new characteristics that would render him *unique*, he lacks any distinguishing features that could constitute a *new* identity. The way we used the term "identity" in this description of McMurphy's condition after his lobotomy is quite common in ordinary English. Talking about identity in this way puts it in the immediate semantic vicinity of terms such as "personality", "individuality" and "uniqueness". Not every person does at any given time possess an identity in this particular sense of individuality as uniqueness of personality. According to this understanding, personal identity is something one has to *achieve*.

Common as it may be, we are not going to adopt the concept of identity in the sense just specified. Instead, in what follows, we will use the term "the personal identity of *x*" solely to refer loosely to the set of conditions whose satisfaction ensures that a person *x* will be identifiable and re-identifiable as *the* person *x*. For, strictly speaking, identity is not an *attribute* of an object, like colour, shape or weight. Rather, attributes can themselves only be ascribed to an object *x* if there are (either explicitly given or implicitly heeded) "criteria of identity" which allow to (synchronically) identify and distinguish *x* numerically from other objects and to (diachronically) *re-identify* (or disqualify) *x* at time *t'* as *the* object *x* already encountered at *t*.²⁶

It should be noted that by speaking of "the personal identity of *x*" in the sense of the set of conditions which (if satisfied) make a person *x* identifiable and re-identifiable as *the* person *x*, the possibility is ruled out that a person's identity could be wiped out leaving behind a person "without identity". McMurphy's personality after having been lobotomised might be utterly dull, but as long as he is not bereft of any capacity that is

²⁶ This is the way "personal identity" is used as a technical term within philosophy since the 17th century. It is noteworthy that it is also the *original* usage of that term – before Locke, personal identity over time was not even conceived of as a problem. From 17th and 18th century "rational psychology" (then a firmly established sub-discipline of philosophy), talk of personal identity made its way first into psychology proper (which as an empirical discipline split from philosophy during the last quarter of the 19th century), and from there, eventually, into everyday life. Interestingly, even the common everyday usage of "identity" in the sense of "individuality" still bears the marks of the original philosophical question about identification and re-identification: When differential psychologists began to describe persons with respect to their location within "spaces" constituted by "dimensions" of personality-traits, *different* persons were (within such a frame of description) distinguishable only insofar as they occupied different places in such a space (i.e. if they differed in at least one personality-dimension). Hence, for all persons to be *identifiable* (within such a frame of description), one needs the assumption that each person has "individuality" in the sense of a *unique* personality.

indispensable for personhood he still is *some* person and as such does possess an identity in the sense that he can be identified and re-identified. However – and *this* is perfectly in line with our use of “personal identity” – there may be concerns about whether the human being we still call “McMurphy” is now in fact a *different* person than the person that underwent lobotomy.

On that basis, the question of whether (old or new) techniques of intervening in the brain endanger personal identity can be exemplified and highlighted by questions like “After that intervention, will there still be *me*?”, “Will the man who’ll come home to our house after the operation still be our beloved daddy?” etc. On the same basis, asking whether techniques of intervening in the brain might also bring about a loss of *individuality* (for instance as a possible long-term effect of socially enforced practices of mental “enhancement”) turns out to be a separate question, which, in fact, will not be addressed in this chapter. However, it is important to realise that the kind of interpersonal comparison required for conclusions regarding the uniformity or diversity of personalities cannot be drawn unless, *beforehand*, the questions of how to identify and re-identify persons as one and the same at different times have been tackled.

The remainder of this chapter consists of four more sections: In Section 5.2, we will address some basic issues and problems regarding the concept of identity and its application to so-called “continuants” in general, i.e. objects that can persist through change – just like persons. Due to a certain abstractness and generality, this section will be the one which (seemingly) is thematically most aloof from the topic of assessing “new techniques for intervening in the brain”. However, it lays the necessary groundwork by providing a general frame for discussing (and resolving) questions regarding the criteria of personhood and personal identity in the later sections. Moreover, it helps to distinguish between the problems that may be dealt with just by clarifying the concepts of identity and of continuants in general and those problems that are actually *specific* to the concepts of person and personal identity.

In Section 5.3 we will systematically analyse the concept of a person, thereby establishing semantic criteria for personhood. This task becomes necessary as soon as the word “person” is understood to mean not just “human being”, and (within the context of this project) this understanding is, in turn, necessitated by the fact that only against such a background concerns and hypotheses regarding possible effects of interventions in the brain on personal identity can be seriously discussed theoretically (and tested empirically). It is important always to bear in mind *this context* because it would be a serious misunderstanding if the considerations in Section 5.3 were interpreted as a venture to establish criteria for use e.g. in medical practice. But, of course, whereas psychiatrists, neurologists or clinical psychologists will certainly have to deal with a variety of psychic disabilities, they will

never – in the context of decisions about treatment – have to face the question of whether their patient is (still) a person or not.²⁷

In Section 5.4, we will discuss critically the pre-eminent philosophical theories about the criteria of personal identity over time, followed by an account of our own. Importantly, this task is different from the one to establish criteria for personhood. It may very well be that certain attributes which must remain unchanged for a person to stay *the* person he or she is do not even enter in the concept of a person. Maybe (just to illustrate the point) a person constitutively must meet certain cognitive criteria, but to remain the person *she is*, must remain stable with respect to some emotive/motivational *character* traits (what and who she cares for, how she is inclined to act in certain circumstances, etc.). For that reason, this section will in fact also have to provide a clarification of the concept of personality and related concepts (like “character”, “temper” etc.).

Finally, in Section 5.5, we will draw some conclusions about the normative questions pertaining to the possibility (in general) that techniques of intervening in the brain might come with side effects regarding personality, personhood and personal identity. Having established criteria of personhood as well as persistence conditions for persons, it will be possible for us to state which *type* of effects would have to count as falling under these categories, and we will also be able to state whether some of these adverse effects would be *subtle* in the sense that they might go unnoticed if not specifically checked upon. With respect to such side effects we will formulate some general recommendations pertaining to research on, as well as the application of techniques for intervening in the brain. Section 5.5 also provides (at its beginning) a summary of Sections 5.2–5.4. We therefore advise readers only looking for information about our general train of thought, the claims we come to maintain and their connection to the recommendations we eventually arrive at (but otherwise lack the time, patience or interest to deal with the rather elaborate network of arguments) to just read Section 5.5. All other readers are welcomed to Section 5.2 now.

²⁷ There could be, however, other such contexts, notably of the type where physicians are commissioned to provide expert testimony in court trials. For assessments concerning a defendant’s “accountability”, “diminished responsibility” or “insanity” are – as will be shown in Section 5.3 – *de facto* assessments of the defendant’s person-status. With respect to treatment decisions, one may wonder whether there is not at least one type of situation where physicians have to deal with the question of their patients’ status as person, namely when they need to know from *whom* to obtain informed consent. However, this – being a matter of legal status – is always legally pre-decided (i.e. the patient either has or has not a legal guardian). In any case, even if the legal status is about to be determined anew by a court, it will usually not be the physician doing the treatment who will get commissioned to provide expert testimony on this subject.

5.2 Identity and Continuants

5.2.1 Numerical Identity

The logical relation of identity is governed by two fundamental principles. The first one is the *principle of substitution*. It goes back to Gottfried Wilhelm Leibniz and says that if $t=s$, then whatever is true of t is also true of s . Therefore “ s ” can be substituted for “ t ” (and, as can be proven, also *vice versa*) in any context where it occurs purely referential.²⁸ In other words: For t and s to be identical they must be indiscernible. Therefore, the principle of substitution is sometimes referred to as the *principle of indiscernibility of identicals*. This principle can be easily confused with a second one, the *principle of the identity of indiscernibles* (*principium identitatis indiscernibilium*), which also has been formulated by Leibniz, and is actually the converse of the principle of substitution. It states that objects which are alike in all aspects *are* identical.²⁹ Taken together, the principle of substitution and the principle of the identity of indiscernibles can be used to define explicitly the logical concept of identity within higher order predicate logic (PL ω):

(=Def.) t (of type τ) is identical to s (of type τ) if and only if for all properties z (of type (τ)) holds: t has z if and only if s has z .

From this definition both our principles follow, as well as *reflexivity*, i.e. that each object t is identical with itself ($t=t$). The fact that it can be explicitly defined in higher order logic is the reason why identity is rightly considered to be a purely *logical* relation.

The logical relation of identity is sometimes also called “*numerical identity*” because, in a process of counting, identical objects count as one. *Two* things – they may be as similar to each other as one pleases – will always differ in at least one property. For example, two otherwise totally alike billiard balls will still occupy different spots and, hence, will be discernible with reference to their spatial properties (that is, their spatial relations to

²⁸ The reason for the clause “where it occurs purely referential” is this: In philosophy of language, the universal validity of the principle of substitution is sometimes questioned with respect to the occurrence of a term “ t ” within so-called *intensional* contexts. Without going into any detail here, we consider this to be a rather unfortunate way of putting things. Far from invalidating the principle of substitution, an analysis of the alleged counterexamples just goes to show that they are not really statements *about t* in the first place and thus would have to be reformulated accordingly in a logically “straight” language. See Hartmann (2003:§23) for a more elaborate discussion.

²⁹ This principle, too, is sometimes rejected. We hold that the metaphysical arguments in favour of its rejection are thoroughly misguided, though. See Lorenz (1969) for an analysis and refutation of the main objections raised against the principle within the historical debate. See also Hartmann (2003:§23) for further details.

other spatial objects). Therefore, the relation of numerical identity is such that each object has it to itself and to nothing else. Accordingly, “*t*” and “*s*” in the formula $t=s$ do not indicate different objects, but just are different linguistic means referring to one and the same object.³⁰ In ordinary language, sentences stating identity can be more than mere tautologies though, since it is not at all obvious in every case that two names or descriptions refer to the same thing, as in Gottlob Frege’s famous example of an informative statement of identity “The morning star is the evening star”.

5.2.2 Identity Through Change: *Perdurantism*

Having expounded the relation of logical identity like we did above, one might raise doubts whether it ever makes sense to claim numerical identity when talking about temporal objects, so called *continuants*³¹. For all such objects – not to mention persons, who from childhood to old age can be expected to be in a more or less constant flux with respect to their “personality” – there is always the possibility of a change in attributes over time. Now it seems that, according to the principle of indiscernibility of identicals, things before and after having undergone any kind of change could no longer count as identical. It is important to see that – for a number of reasons – we can’t just bite this bullet, simply give in and, consequently, restrict the notion of identity to “timeless, unchanging objects” like, e.g., numbers. Such a conception would be at odds with the most fundamental presuppositions upon which the very frame of our languages (natural and formal alike) is based: By using names we subscribe to the conviction that things can be re-identified in various situations at different times as the *same*. If there was no object permanence, the use of a name more than once could never be correct. Furthermore, if there was no chance to recognise objects as one and the same, all natural laws/equations which functionally describe a change in

³⁰ As objects are counted by identity, it is often remarked (and rightly so) that certain natural-language formulations of the principles governing identity are unfortunate insofar as they seem to presuppose that the “objects” the identity of which is stated are in fact “two”. The most obvious slip would be “If two objects *t* and *s* are identical, then ...”, which in the antecedent clause assumes a manifest contradiction. And in one of the formulations of the principle of substitution we cited above, we wrote: “For *t* and *s* to be identical, *they* must be indiscernible.” However, inconveniences like this can easily be avoided in a more canonical mode of speech, in which we may avoid using anaphoric cross-references or other constructions grammatically demanding plural form. For instance, “if *t* is identical to *s*, then *t* is indiscernible from *s*” would be perfectly fine.

³¹ This term has been coined by W. E. Johnson (1924) because it is unburdened by the problematic connotations of the term “substance”. He introduced it against the backdrop of the classical metaphysical dichotomy of “event” and “substance”, indicating that continuants are objects which persist through change, whereas events do not change, but rather consist of changes.

state of given objects over time – like the law of free fall – would be false.³² “Strictly speaking”, objects would never change, but rather continuously pop into and out of existence. For any practical purpose it would, obviously, be impossible to handle the infinite explosion of the number of objects which abandoning the idea of object permanence would entail.

The way out of this predicament has been shown by David Lewis (and others – but see especially Lewis 1983: “Survival and Identity”). The basic assumption is that, as soon as one is talking about a continuant, any predicate *Q* expressing a property liable to change is implicitly relating the object to a point in time, such that the *same* object can be *Q* at one time *t* and *not Q* at another time *t'* without violating the principle of substitution. As a matter of coherence, *spatial* objects have therefore to be regarded as extended not only in space, but also in time: they are “4-dimensional”. Or, in more general terms, all continuants (e.g. events) are *at least* extended in time, and spatial continuants are extended in *both* space and time. That entails that continuants have temporal parts, so called *stages*.³³

³² This raises the question whether object permanence is a condition of the possibility of scientific experience or even of experience in general. That this is indeed the case is argued for by Immanuel Kant in his “Kritik der reinen Vernunft” (1904/11 [AA III]: B 178). We will have to put the presentation of the argument on hold, however, until Section 5.3, where it will be part of our analysis of what is entailed by the assumption that some being is a person.

³³ Antagonistic to the position we adopt here is the view that continuants are at all times “wholly present” (see e.g. Zimmerman 1998). Often, the two competing positions are terminologically highlighted by saying that according to the “4-dimensionalists” continuants *perdure*, whereas according to the “wholly-presentists” they *endure*. Sometimes, the very term “continuant” is reserved for the endurantists – which then, of course, implies that the perdurantists *deny* the existence of continuants. This, however, is unfortunate as many perdurantists – like e.g. Lewis – explicitly use the term “continuant” for the 4-dimensional objects they conceive of. Anyway, we fail to see how the endurantist can solve the problem of continuants as changing and at the same time “enduring” without committing a violation of Leibniz’s principles, which we saw are logical truths and therefore not open to revision. Maybe the best way to approach the endurantist’s position is to lay open the intuition motivating his rejection of “4-dimensionalism” and then show how this intuition, however sound, is employed incorrectly. The sound intuition is that, in the world we live in, we are able to perceive continuants. However (so it is argued), if a continuant was a perduring object, we could never perceive or otherwise have contact with it – all we could perceive or have contact with would be temporal stages of it. Hence, to perceive a continuant, it must be “fully present” at all times. What is wrong with this argument is that, according to the same logic, we would also never perceive an *enduring* spatial object, as, for instance, we always only see some very small spatial part of it, namely the surface-part which is presented to us. And from this, we could then (again by the same line of reasoning) infer that, as we really *do* see spatial objects, they must be “wholly subsistent” in all their visible parts. But of course this is all but a conceptual confusion. To dispel it, we have to distinguish between two ways of using the expression “to see”: one use according to which we always see₁ just certain parts of an object, but never the whole object, and a second use according to which we always see₂ objects simpliciter. And the relation between these two usages is very simple: whenever one does see₁ an arbitrary surface-part of an object, one does see₂ that object. And along the very same lines, one does of course see₂ it, if one does see₁ a surface-part of one of its stages.

However, this concept of continuants is still not sufficiently elaborate. For unless one puts further restrictions on the identity of objects over time, continuants would have to be thought of as *eternal* inasmuch as no change (short of “complete vanishing”) would ever end their existence. The first step in resolving this further puzzle is to distinguish between two ways of talking about “existence”. One kind of sentences stating the existence of something can be analysed using the existential quantifier of predicate logic. Only because of their superficial grammatical form, sentences in natural languages like “Lipid-soluble toxic substances exist” seem to attribute a property (“existence”) to objects. But their logical form is revealed by reading them “For at least one x holds that ...”. In a second sense, however, the verb “to exist” is indeed used as a predicate proper. In sentences of the respective kind, it can be exchanged by predicates such as “alive”, “enduring”, “present”, etc. So, if someone (implausibly) stated “Adolf Hitler still *exists*” he would thereby claim that Hitler was still *alive*.

Sometimes both ways of talking about existence are used together. Take, for example, the sentence, “There is a painting of Titian that was destroyed in a fire and so no longer exists today.” In analysing the formal structure of this sentence one needs to use the existential quantifier. If this would imply attributing a property, then the first and the last part of the sentence would contradict each other. Actually, the existential quantifier could not at all be employed to formalise statements about things long gone if it stood in for a predicate. But its function is rather to indicate the *reference* of a statement. We can understand what someone uttering the above sentence is talking about, no matter whether any painting of Titian ever was destroyed in a fire. On the one hand everything we can at all talk about *exists* “timelessly” in the sense that the semantic relation of reference is itself timeless. On the other hand we can talk – in the second sense specified above – about the existence of continuants over certain time-spans. In order to avoid confusion we shall use the verb “to persist” when we will be talking in the “non-timeless” or attributive sense of existence. For instance, from the statement, “Napoleon is to be considered the greatest military strategist of all times” it follows that there is someone who is to be considered the greatest strategist *now*, but not that this someone does also *persist* now.

As said above, criteria of identity of objects over time are needed to allow the possibility that continuants can actually cease to persist. Furthermore, such criteria are also required to decide, in the first place, whether any given two “stages” at different times are stages of the same continuant or not. And here we are again homing in, in a most general form, on the main topic of this chapter, the question whether “snapshots” of a human being, taken before and after an intervention, might not belong to the same person. A few quite fundamental philosophical remarks might help the reader to get a better understanding of that question.

First of all, objects do not exist “in themselves”. They are constituted by our distinguishing them through means of general concepts. In an impor-

tant sense, language actually constitutes our world. Certainly, not all the linguistic means I might use to describe an object are of equal importance with respect to its constitution. I might describe the thing standing in front of me as “relatively big”, “stained”, “ugly”, as “a present from my mother in law” and as “made of porcelain”, but still you would not know really what I am talking about. It would have been of great help to you if I had told you in the first place that I am talking about “the cup” standing in front of me. In a way, all the other properties I attributed to that cup by applying those various descriptive expressions to it happen to be merely “accidental”. The cup could be rather small, clean and pretty, and it would still remain a cup. Its being-a-cup, on the other hand, one might call its “essential property”. This dichotomy has been established by Aristotle, who also referred to an object’s essential property as to its “form”. But we would like to avoid these expressions as we do not want to endorse *ontological essentialism*. Without going into details, the distinctive difference between our way of thinking compared to Aristotle’s essentialism is that we consider the “essentiality” of a property as *being relative to a description*, which in turn will, generally, be justified only in relation to its “fit” with respect to the goals pursued within the discourse in question. To emphasise the context- and language-dependence of our approach we will talk about *constitutive predicates*, rather than about essential properties.

Having said this, we hope it is still comprehensible for us to maintain that, for any given object distinguished by us as such, there will be a constitutive predicate. Since every object is constituted by the application of a constitutive predicate, a necessary condition for any object to persist (rather than to perish or to be transformed into another object) is that this predicate “keeps on applying”. If we stick to the example given above, and consider the predicate “cup” as constitutive for a certain thing, then it may be painted differently or lose its handle and still remain “the” cup. However, if it breaks, what remains is not a cup anymore and so that particular continuant has ceased to persist.

5.2.3 Material Identity and Material Continuity

Introducing the concept of constitutive predicates into logic brings another noteworthy consequence: Unless we adopt it, there is for any *P* which is also *Q* always an *identical Q* which is also *P*. This is no longer the case if we bring constitutive predicates into play to avoid having all objects existing (persisting) eternally. Certainly, if we are faced with a constituted *P* that happens to be *Q*, we can still say that there’s also a constituted *Q* which is, accidentally, *P*. However, on logical grounds of the substitution principle alone, “the *P*” need not be identical to “the *Q*”, namely, if there is a time where the *P* does still/already persist while the *Q* has ceased/not yet begun to do so (and *vice versa*). Imagine, for instance, a little boy who moulded a toy dog from plasticine. As long as he is playing with the dog he will typically treat it as a dog,

being totally oblivious to the fact that it is made of plasticine. Let's say, after having done with it, the boy tosses the toy dog into some box. Then, after some time, he might feel again like playing with plasticine. When eventually he finds the discarded plasticine dog he may treat it as nothing but a blob of plasticine, quickly squeezing it beyond recognition. – This innocent example illustrates how what is taken to be constitutive of an object as against to what is deemed to be accidental can change with the attitude that is taken towards that object. Furthermore, it becomes clear that the continuant we referred to as the “toy dog” is not identical to the plasticine it consists of, since the latter is there before the former comes into persistence and after it ceased to persist.

This last insight can be generalised: *No material object will be identical to the matter it consists of.* First of all, this obviously holds true for all those things of which one can remove or replace some material part, or which can lose or exchange some amount of the matter they consist of, without thereby changing their identity. Actually, there are not many things whose identity is constituted rigorously by the exact amount of matter they consist of, to the extent that they would not count as the same when some tiny fraction was removed. A notable exception could be the prototype for introducing the international mass unit kilogram – a cylinder made of platinum-iridium-alloy stored in the “Bureau International des Poids et Mesures” in Sèvres near Paris. In this case one could argue that the prototype's identity is actually changed if you scrape off just the tiniest part of it. But even then, the continuant we refer to by the definite description “the prototype for the kilogram” is *not* identical to the noble metal it consists of, since – once again – the latter clearly was there before the former was introduced in a meticulously controlled production process.

By arguing that “the *P*”, in general, is not identical to “the *Q*” it consists of, we refute the view that for spatiotemporal continuants identity as such is constituted by *material identity*. To further illustrate this point, we will now consider the persistence conditions for human bodies, for this will prepare the subsequent discussion of personal identity. Regarding the identity of a human body, two of the hard questions one can ask are: When exactly does it come into being? And when exactly does it perish? Fortunately enough, these are questions we need not answer here, but at least we would plead for a pragmatic way of approaching them. In our opinion, there is not just one *true* answer to each of them. It rather depends on practical purposes what answer can be taken to be *appropriate* in a certain context. Long before the arrival of modern imaging techniques, which today provide us with various insights into a mother's womb, it was a known fact that human bodies do come into persistence *before* they are born. Even today, though, the difficulty is to tell *at what time* in between conception and birth they come into being. On the one hand, thinking of the continuity of this development, it seems quite arbitrary to pick out any particular stage of embryonic development as

a “starting point” for the body. On the other hand, however, it seems rather odd to choose the fertilised egg as a starting point, since it bears absolutely no similarity to the human body that later on we identify by its traits as one and the same in spite of all its changes. We find ourselves in a similar kind of quandary when it comes to the exact stage of decomposition from which onwards it is no longer appropriate to call the mortal remains of a human a human body. The important point is that, apart from particular practical contexts for which the answer to this kind of questions would actually make a difference, we do not have any criteria to judge the “absolute” appropriateness of different possible answers.

So, considering the human body as a continuant, it is not easy to specify a starting and an end point for this 4-dimensional entity. However, no matter how this problem is settled, one will find support for the assumption that the relationship between any particular human body and the matter it consists of is not one of identity. Much to the contrary, regarding its material substrate the human body offers a striking example for identity through change. Assuming the (living) body’s material substrate means its chemical composition, it is well known that the body is in permanent metabolic exchange with its environment via assimilation of nutrients, excretion, etc. Consequently, there probably is not a single molecule – or cell for that matter – that is part of a particular body from its elusive start till its elusive end. However, the example of the human body also indicates that it would be premature to move to the opposite extreme by claiming that the identity of material objects wasn’t constrained in any way by conditions regarding their material substrates. For, regardless of the human body being in a permanent reconstruction process, we seem to require some sort of material *continuity* between two body stages if we are to accept that both of them belong to one and the same body. To grasp this point, just imagine teletransportation was actually possible in the following way³⁴: In one place all the information about a human body is encoded and then transmitted to another location

³⁴ Readers who are not familiar with what is called “thought experiments” in philosophy may be somewhat alienated by what seems to be a lengthy discussion, not of empirically relevant facts or possibilities, but of far-fetched science-fiction scenarios. However, it is important for the reader to understand that the reason for the regular occurrence of such examples within philosophy is *not* that philosophers take a special, escapist liking to science fiction or weird things in general, which would have no relation to, or bearing on the urgent problems we face in real life. Rather, such discussions of counterfactual situations are part of the philosopher’s genuine task of analysing and clarifying the meaning of concepts. Thought experiments invoking radically counterfactual (but logically consistent) situations are a valuable tool for elucidating the meaning of a concept (including its links to other concepts) because they allow testing of its *limits*. With respect to any proposal that the meaning of a concept should be taken to consist in “such-and-so”, thought experiments will test whether that proposal is “robust” with regard to the entire range of possible experience with which future developments may or may not confront us, or whether it implicitly draws on the stability of contingent facts.

where it is used to create an exact copy of the original body, which itself is destroyed in the process. Despite its exact similarity we cannot admit that the copy-stage belongs to the original body. For imagine the last step of destroying the original body was omitted: Then we were faced with two exactly similar human bodies, the stages of only one would be counted as belonging to the body prior to the process of teletransportation – the one which bears the relation of material continuity with the body prior to teletransportation. Thus, in “teletransportation” the copy is nothing but a copy, regardless of whether the original body is destroyed or not.³⁵

One may still think there is something puzzling about material continuity as a necessary³⁶ criterion of identity for human bodies: Take on the one hand two stages of one and the same human body separated by a long stretch of time and compare the relationship they bear to each other with the relationship between one stage of a human body immediately prior to teletransportation and one stage of this body’s copy immediately subsequent to that process. According to what we stated above, in both cases you might not find a single molecule or cell that establishes a direct material connection between the two respective body stages. What is more, in the first case the two stages look very dissimilar, while in the latter case they look strikingly similar. But still it seems plausible to maintain that only in the first case the two stages belong to the same human body. In order to come to terms with this “paradox” it is helpful to compare the relation between two stages of a human body, one at the beginning and one at the end of its life, with two parts of a rope at opposite ends: Even though not a single fibre runs through all the rope, so that there is no immediate connection in this sense between two parts at opposite ends, they still are obviously connected. In much the same way there is a material continuity between stages of a human body at young and old age, respectively, that is lacking between the body and its “teletransported” copy.

Quite obviously, the criterion of material continuity – allowing for persistence despite loss/exchange of matter – can be generalised to cover not

³⁵ Teletransportation violates the criterion of material continuity insofar as the “rebuilt” material parts of a “transported” object cannot, not even in principle, be considered as numerically identical to the destroyed parts. The reason is that material continuity presupposes *spatiotemporal* continuity. It follows that for “teletransportation” to be taken serious, its “theory” would have to be quite different from the one involving the trias “scanning”/“destroying”/“rebuilding”. It would have to assume that (under specified circumstances) objects could pass through a fourth *genuinely spatial* dimension that directly connects regions which are only indirectly connected in three-dimensional space (for illustration, think of a paper that is folded up onto itself).

³⁶ Material continuity is at the same time obviously not a *sufficient* condition for the identity of the human body. For this, the constitutive predicate “human body” must remain applicable, too. (If it decomposed or transformed into another type of body, it would have ceased to persist, regardless of maintained material continuity.)

only human bodies, but at least all living bodies, i.e. organisms. Can it be generalised to cover *all* bodies? Many philosophers (mostly following Locke and Nidditch 1975:II, 27, § 3) think there is a difference between organisms and inorganic bodies in that the criterion of material continuity only applies with respect to organic bodies, while an analogous (necessary) criterion of identity over time for inorganic bodies would be material identity.³⁷ Though it is obvious that the latter criterion would eventually have to be relaxed (otherwise the removal or exchange of the tiniest part would again change the identity of inorganic bodies), the overall proposal has indeed a certain plausibility to it: Assume, for instance, my wife had once bought me a tripartite lamp consisting of a frame, a shade and a bulb. Over the years, I exchanged the bulb, then the shade, and eventually the frame. It would ring rather peculiar if I nevertheless would insist that what I've got there hanging in my room is still the lamp my wife once bought me – even if in exchanging parts I had kept to the same brand and model. For if I had given the original parts to a friend who eventually put them together again, *his* would be considered to be the lamp my wife had once bought me.

To put to test the hypothesis regarding the range of application of the criteria of material continuity and material identity, let's see how they fare with respect to the famous puzzle of the "ship of Theseus". This puzzle, which traces back to Plutarch, runs as follows: Over time, worn planks of the ship of Theseus are replaced, bit by bit, by new ones, so that eventually none of the original planks remain in it. Meanwhile, however, the worn planks are used to build a "new" ship by assembling them in just the way they were arranged in the original ship. After completion of this reconstruction process, there are two ships looking pretty much alike except that one is made of new planks (ship 2) while the other is made of worn planks (ship 3). And now the question is *which* of them is identical to Theseus' original ship (ship 1).³⁸ This being a good puzzle, no answer is completely satisfying. On the one hand, only ship 3 satisfies the requirement of *material identity* as it is made of the very same parts that constituted Theseus' original ship – thereby making a good case for claiming that this is indeed the same ship. On the other hand, replacing a worn plank is obviously not enough to end the persistence of a ship and create a new one instead. But if that is so, then persistence must be preserved through *each single such step* and, hence, ship 2 must eventually inherit identity with ship 1. Now, if we ask what it is that connects ship 2 (but *not* ship 3!) to ship 1, we find that it is in fact the relation of *mate-*

³⁷ It is clear that, insofar as "material identity" means "consisting of the same material", one will eventually have to face the task to define the criteria of identity for *that*.

³⁸ More precisely, the question is which of these stages (2 or 3) is a stage of the ship which stage 1 belonged to (as the different stages cannot be identical anyway). However, for the sake of convenience, in the following we will use this precise formulation only if otherwise there would ensue pseudo-problems.

rial continuity – which seems to make a strong case here even if Theseus’ ship is an inanimate object.

While searching for the one true answer to this puzzle will give the metaphysician a serious headache, we once again opt for context-specific solutions. For example, a practical need for a straight answer to the question as to whether ship 2 or ship 3 is identical to ship 1 could be constituted by a legal interest to settle the question of ownership. If, for instance, Theseus sold the replaced pieces of his ship to a scavenger who then came up with the idea of reconstructing her, then it is blatantly clear that Theseus is the owner of ship 2, but not of ship 3 – and thus ship 2, but not ship 3, would be regarded as identical to ship 1 in this context. To see the problem from another context, consider the following clever modification of the Theseus puzzle, borrowed from S. Marc Cohen³⁹: This time Theseus’ ship is not replaced during its journeys, but, having been put out of service, displayed in a museum. Suppose now some thieves are trying to steal the ship by removing its pieces one at a time. In the museum, the stolen pieces are replaced with look-alikes while they are secretly reassembled elsewhere in the original way. In this modified case it seems quite clear that only the ship in the thieves’ hideout may be considered to be identical to the original ship whereas the ship finally left in the museum is a fake. Even though our legal intuitions point into opposite directions in the two cases discussed here, the respective judgments seem fairly obvious. The lesson is that questions regarding identity that are bewildering as long as they are posed “out of context” frequently may be settled quite agreeably in the light of concrete practical purposes. We will have to keep this lesson in mind when it comes to the real, serious concerns pertaining to personal identity associated with the medical procedures that are the principal topic of this study.

To avoid narrowing the scope for finding practical solutions for identity issues, we deliberately refrain from setting very strict criteria for persistence through change. For instance, the way we determined the concepts of identity and persistence of continuants in general still *leaves open* the logical possibility that at least some kinds of things can cease to persist and then later begin to persist again. Consider once again the example of the plasticine dog. Say, while your child is playing with that dog you inadvertently step on it, squashing it completely and leaving him crying his heart out. To console him you take up the task of moulding “the same” dog until at least your boy is convinced that “here it is again”. It is not that important whether *you* would actually accept this as a genuine case of “resurrection” or rather think by yourself, “well, it is a different dog looking more or less similar and made of the same material”. In every individual context we may have good reasons to

³⁹ “Identity, Persistence, and the Ship of Theseus”, <http://faculty.washington.edu/smcohen/320/theseus.html>, accessed on December 8th, 2006.

treat such cases one way or the other⁴⁰. So it turns out as an advantage rather than a shortcoming of our account that resurrection is not a priori ruled out for *conceptual* reasons. If required, constraints of “unbroken continuity” could easily be introduced to further regulate the identity of certain types of objects – like we suggested in the case of organisms.

Another theoretical advantage of our account is that it does not enforce an affirmative answer to the question whether differences in *modal properties* rule out identity. Modal predicates such as “is *necessarily* mistaken” or “is *possibly* harmful” are sometimes considered to ascribe to objects a special kind of properties called “modal properties”. If one subscribes to this (*de re*) conception of modal expressions it seems natural to maintain that – according to Leibniz’s laws – objects differing in nothing but their modal properties must nevertheless be different. In our view, however, modal statements are *not* statements about objects, but statements about statements (*de dicto*). The reason is that the *de-re* view together with the principle of substitution must inevitably lead to absurdities. For example, if the statement, “It is necessary that $9 > 5$ ” is taken as ascribing a “modal property” to 9, then from its combination with the true sentence, “The number of planets of our solar system = 9” one could infer via the substitution principle: “It is necessary that the number of planets of our solar system > 5 ”. While we still may talk in a derivative sense about the modal properties of *objects*, we always have to keep in mind that an object does not have its “modal properties” *independent of the expressions used to refer to it*. Accordingly, differences in modal properties do not *per se* exclude identity.

The motif behind the inclusion of modal properties by some philosophers is the fear that, otherwise, different objects might not be discerned, for example a cup might falsely be considered identical to the porcelain it consists of, in cases where both would come into existence and vanish simultaneously. However, to us it appears that, apart from a possibly different time-span of persistence, there are always other *actual* differences to be found: The Rosenthal-cup, for example, is worth 20 Euro, but the “mere material worth” of the porcelain is much less (think of the longstanding debate regarding customs duty on CDs or diskettes).

5.2.4 *Perdurantism Refined and Defended*

Let’s now return to the main question of how the riddles surrounding the concept of identity through change can be resolved. Our first step was to introduce the concept of a continuant with stages as temporal parts. This led

⁴⁰ Maybe the best example for a field where talking of “resurrection” is the common mode of speech is the context of historical buildings. During its history, a bridge, a castle or a city hall may have been not only enlarged, restored or relocated, but even destroyed and later rebuilt. It is also quite obvious that it would be futile to search for definitive ontological criteria with respect to the question of whether buildings “really” persist through such-and-such changes or not.

to a reconciliation of identity (in the strict logical sense) with change. The second step was the introduction of constitutive predicates, which allowed for continuants to cease to persist. Above we said that, for a continuant to persist, at the very least the constitutive predicate has to “keep on applying”. It is now time to reveal that this, if taken *a la lettre*, is either circular or not quite penetrating to the fundamentals. This can be easily seen if we ask *to what* the constitutive predicate is supposed to “keep on applying”. Obviously, if we speak in those terms, we presuppose *having already constituted* some “subsisting” continuant. For example, if we would agree upon material continuity (implying spatiotemporal continuity) as a necessary and sufficient criterion for *bodies*, we could then say that “this cup” persists as long as the constitutive predicate “cup” applies to “this body”. But even though it is indeed true that bodies are, in a certain way, the primary continuants⁴¹, neither can all continuants be thought of as constituted through bodies in this way (e.g. events cannot), nor do the persistence conditions for all sorts of continuants presuppose that some other, subsisting continuant must persist at least as long (e.g. for a traffic congestion to persist there does not have to be one individual car involved in it *all* the time). Fortunately, there is a non-circular, even if somewhat roundabout way to formulate our criterion of the “continuously applying constitutive predicate”: With respect to any given continuant-stage that exemplifies some continuant through a constitutive predicate (e.g. “this cup”), only such later continuant-stages to which at least the constitutive predicate (i.e. “cup”) applies can be considered as possible continuations of *that same* continuant. From this formulation it immediately becomes clear that there must be stronger criteria for really fixing the identity of continuants over time. For many (though, as we will learn from the example of persons, not for all) embodied continuants, these criteria often conceptually involve the identity of the subsisting bodies over time, and thus their spatiotemporal continuity (which is implied both by the criterion of material continuity and by the criterion of material identity).

However, even in conjunction with continuous application of the constitutive predicate these criteria often do not suffice. Imagine, for example, I start out with an 80386 IBM computer which I gradually change over the years, replacing the motherboard, the case, the monitor, the mouse etc., eventually ending up with a Pentium IV with flat screen. In this situation, the constitutive predicate “computer” obviously still applies to the object I am working with. But is it really still the “same computer”, just very fancily changed? Maybe we wouldn’t want to say so and accordingly would need to agree on stronger criteria for “persistence of computer identity”. Maybe, insofar as the central processing unit (CPU) can be considered to be a computer’s “brain”, one would

⁴¹ They constitute space and thus our universal frame of reference, in relation to which all other continuants get located directly (like lightning) or indirectly (like a smile or an itch).

want to argue that after replacing the CPU with one of a different type a computer is no longer the same (in our example, a 80386 is turned into a Pentium IV), while changes at the periphery (mouse, monitor etc.) do not interfere with a computer's identity. Once again, this prosaic example seems fit to bolster our general point of view that questions regarding identity should be handled as pragmatic rather than as "metaphysical" issues. But what's even more important is that the example also shows that, in general, what is needed in addition to a continuously applying constitutive predicate *P* are criteria stating which *further* predicates *Q* must hold for any specific *P* to stay the "same *P*".

In the next section we will turn to the task of setting out what it means if the constitutive predicate "person" applies to a continuant. Thereafter we will propose criteria that further specify what mustn't change for a person to stay the "same" person. But before doing so, we would like to discuss a few more theoretical issues concerning identity of continuants in general that might appear somewhat idle to non-philosophers, but play an important role in the technical literature. One such issue, which some presume to offer an objection to four-dimensionalism, is typically addressed by the keyword "overpopulation". It is usually exposed by talking about persons, but we prefer to address it by talking about "cups" so as to emphasise its relevance for the notion of identity of continuants in general.

The main strand of arguments referring to overpopulation relates to the problem of individuating continuants. If you were a philosopher sitting in front of a cup you might wonder how many cups you are actually facing. Is it just one cup? Or maybe two cups, one with a handle and then the same cup again, but this time counted without the handle? Or are there even more of them? Concerns of that kind can be resolved by stipulating that continuants are "maximal" in space. So, for situations in which *n* *P*s are to be counted, if a certain *P* (say *a*) is counted once, no "further" *P* is to be counted which is constituted of just part of the matter that makes up *a*.⁴² Similarly, but less trivially, continuants are maximal with respect to time as well. That is, no temporal stage of a cup is *another* cup.

Thought experiments involving amoeba-like "splitting" also give rise to concerns regarding the concept of a continuant. Actually, the case of amoebas itself can be used to discuss the issue (the advantage being that this is *not* a mere thought experiment): Suppose, an amoeba is dividing, with an end result of two alike but numerically different amoebas. This is what in the literature is called a case of "fission"⁴³, in which we are confronted with three

⁴² As always, we must not forget that criteria like this must be understood as pragmatic, not as metaphysical truths. There are contexts where we might want to use criteria of counting that disregard maximality. For example, we might prefer to stick with the habit of saying that *two* animals are shown in the famous ambiguous duck/hare-picture.

⁴³ Usually exemplified by thought experiments using "duplicators" or "teletransportation going wrong".

things: one previous to the fission point (amoeba 1), and two of them in different places subsequent to it (amoebas 2 & 2'), indiscernible from each other except for their spatial properties. Certainly, this indistinguishability may be limited to the time frame immediately before and after the fission point – for instance, one of the two resultant amoebas might get injured shortly after the fission. But still, other than in the case of the ship of Theseus there is (at least *ex hypothesi*) no difference whatsoever regarding the material properties of amoeba 2 and amoeba 2'. Therefore, the question as to which of them is “identical to” past amoeba 1 seems even more intractable. One might argue as follows here: “In the case of fission, amoeba 2 and amoeba 2' can't be both identical with amoeba 1, because then they would (per transitivity of identity) also be identical to one another.⁴⁴ That would mean neither amoeba 2 nor amoeba 2' should be regarded as identical with amoeba 1. Amoeba 2 and amoeba 2' are “new” whereas amoeba 1 simply ceased to persist. But that, in turn, must eventually lead us to the generalised conclusion that *never* should *any* object be taken to be identical to some earlier object, as persistence clearly shouldn't depend on the totally *accidental* (“external”) issue of whether (in the meantime) a fission has *de facto* occurred or not.” – There must be something wrong with this train of thought since it implies that no object can survive the shortest stretch of time, a notion already found untenable above.

To solve the problem, the so-called “best candidate theory” was proposed (see e.g. Nozick 1981). According to this view, in the case of fission one has to pick the object which is the “best candidate” with respect to the claim of being identical with the object that was there before the fission occurred. For example, if teletransportation fails in the sense that, while a copy of the object actually appears elsewhere, the object is not destroyed at the original location, the object at the original location would be the “best candidate” because it alone bears the relations of material and spatio-temporal continuity with the object before the fission point. Only if no “best candidate” can be found, the object prior to the fission point has ceased to persist (as it has been replaced by at least two “new” objects). It has to be conceded that the “best candidate” theory contains some grain of truth, but only insofar as there indeed have to be independent criteria for deciding whether two stages belong to one-and-the-same continuant or not. If some object stage S_i does not pass the test with respect to another object stage $Z_{t-\Delta t}$ while another

⁴⁴ This argument can be couched in terms of fusion, too. To avoid the realm of thought experiments, one has to draw on examples of symbiotic or parasitic fusion (e.g. the case of sperm/ovum, the hypothetical development of mitochondria etc.). Sperm and ovum cannot both be counted as identical with the “fused” zygote, for then they would have to be identical to one another, too – which is not the case. However, the argument from fusion does not yield any surprising or unsettling results as, according to common intuition, things are usually not expected to “survive” fusion anyway.

“candidate” S'_t does, then S'_b , but not S_t is to be regarded as a continuation of $Z_{t-\Delta t}$. Still, the “best candidate” theory fails insofar as it entails more than that, namely that in fission there *either* is one “best” candidate, which then will count as the continuation of the original object, *or* there are at least two equally good candidates, in which case the original object ceased to persist. A “fission type” thought experiment, originally devised by Wiggins (1967) and since then often recycled and modified by others (e.g. Parfit 1984:ch. 12, sec. 89), can be used to show that this leads to (conceptually) absurd consequences: Imagine brain transplantation would be possible and a (most probably mad) neurosurgeon was about to transplant the two hemispheres of your brain into different new bodies. Let’s assume, for the sake of argument, that the new bodies are exact copies of your old body (Parfit invites us to assume they are in fact the bodies of your two brain-dead triplet brothers), and also that both hemispheres are (maybe deficient, but) equally good candidates for being continuations of your brain as it was before.⁴⁵ Now, as in such a case the “best candidate” theory *ex hypothesi* decrees that the original person (“you”) will cease to persist, it would be very rational for you to bribe a nurse to destroy one of your brain-hemispheres so that only the other gets transplanted (in this vein e.g. Noonan ²2003:ch. 12.5). For in *this* case, there would remain a “best candidate” and you would survive! It is very obvious from this that the verdicts of “best candidate” theory are too arbitrary to be of help in solving fission-related problems regarding the concept of “continuants”. Some better solution has to be found.⁴⁶

Luckily, we do not have to look very far. Our diagnosis is that fission engenders problems only if it is addressed by the wrong question: As objects *themselves* are extended in time, the seemingly innocent question, “Is object

⁴⁵ This is allegedly true for at least some individuals where none of the usual functional asymmetries between hemispheres is to be found. Functional asymmetries tend to be strongest in right-handed males and much less in females and left-handed people.

⁴⁶ Holding on to the “best candidate” approach, Derek Parfit (1984) takes these considerations to show that “*identity is not what matters in survival*”. Of course, this view is extremely counterintuitive, and not *just* because, taken literally, it is incoherent: If it’s not *me* at a later stage, then I didn’t *survive*. What Parfit means, however, is that “division” thought experiments can be used to show that what we care for is not that *we* will still be there later, but rather that there are future persons (at least one) which are, in a sense, *continuers* of us. For – as, *in truth*, we would find the idea of bribing the neurosurgeon into destroying one of our brain hemispheres extremely inappropriate – what *we* are interested in *cannot* be our literal survival. It is easy to see from this that the whole argument *presupposes* best candidate theory to be correct (this has been shown in all desirable clarity by Harold W. Noonan [²2003:ch. 9]). So, in the face of the undisputable fact that the assumption of a lack of interest on our part in *our* survival is extremely counter-intuitive, this very consequence of “division” thought experiments rather seems to show the absurdity of the “best candidate” approach once more. In any case, if there are rival theories regarding the persistence of continuants which do *not* imply that we lack interest in our survival, they clearly should be preferred.

O1 at t_1 identical with object *O2* at t_2 ”, does not properly confer the meaning that is commonly intended. To consider an object at a certain time means to consider a *stage* of that object. However, stages at *different times* are never identical anyway, so the question in question would, strictly speaking, *always* have to be answered to the negative! However, it makes perfect sense to ask “Is *O1* identical to *O2*?”, or “Are *O1* at t_1 and *O2* at t_2 both stages of the *same* object?” Here, the negative answer is not obligatory. It only is if objects *O1* and *O2* are discernible at (at least) *some* point in time t – and this is how it should be. Such a distinction between objects *O1* and *O2* would have to be registered, for instance, if there is any point in time for which their spatial location differs.⁴⁷

The decisive point, however, which still has to be made, is this: Regarding the correct questions, there is the (logical) possibility that amoeba 1 and amoeba 2 are stages of the same object and *also* that amoeba 1 and amoeba 2’ are stages of the same object. This does *not* imply that amoeba 2 and amoeba 2’, too, are stages of the same object!⁴⁸ (In fact – due to their different spatial locations after the fission point – we *know* that they are not). Generally put: Two different objects (that is, objects which aren’t identical) may nevertheless *share stages before or after a certain point in time*. And, of course, these points will be fission or fusion points.

But, while this clearly “defuses” the threat that fission presents to the concept of continuants and their persistence in time (thereby also doing away with the need to resort to “best candidate” theory⁴⁹), we now seem to be faced with the problem of “overpopulation” again, in the sense that where a fission occurs there apparently have been *two* objects before the fission point even though we thought there was only one. But, of course, before the fission point in question is actually reached, it makes no sense to distinguish two things (as, before that point in time, they are literally indistinguishable). So, in our example, before

⁴⁷ Of course, any spatial object is such that spatial parts of it do occupy different regions in space. So, if one feels inclined to ask why a spatiotemporal object cannot occupy two *distinct* spaces at the same time, the question needs to relate to the object as a *whole*. Under this reading, the assumption that an object occupies two distinct spaces at the same time would bring about logical and categorical mishaps: Suppose, for example, there were two spatially distinct “occurrences” of “one-and-the-same” marble, then the resultant discontinuous object taken as a whole could no longer be called a marble at all.

⁴⁸ This is in sharp contrast to the results yielded by the wrongly put question in which stages were confused with the objects they belong to. By mistakenly asking which “later” objects are *identical* to some “earlier” one, one is unavoidably led to the problem of transitivity of identity.

⁴⁹ In this way we also avoid the nasty consequence of that theory, namely that our literal survival does (“in truth”) not matter to us. For, in cases of real fission, *no* one ceases to persist – and *this* is the reason why, under the (idealised) circumstances of Wiggins’ thought experiment, the idea of bribing the mad neurosurgeon into destroying one of our brain hemispheres before the transplantation doesn’t appeal to us (and *not* some alleged indifference with respect to our literal survival).

the amoeba has divided there *are no* two amoebas. To better understand how this is still in step with our view that different objects could share stages before a fission or after a fusion point, it is helpful here to distinguish between two kinds of situations in which to use the expression “before”. It is one thing to talk about events happening before a time t^* while t^* has not yet arrived, and quite another to talk about what was before t^* when t^* has passed. After the fission we might think differently about amoeba 1, with which we were acquainted before. Alluding to Arthur C. Danto (1985: especially ch. VIII), one might say that, in a way, the future can change the past.⁵⁰ However, it may come as a relief to those who feel uneasy with the last statements that even after t^* we are not necessarily obliged to say that “two” things (amoeba 2 and amoeba 2’) existed before t^* . In order to come to terms with cases of fission and fusion, David Lewis (1983: “Survival and Identity”) introduced the notion of *tensed identity*. According to his proposal, things are to be counted by *identity-at-t* instead by identity simpliciter⁵¹: *a* and *b* are identical at *t* if they are indistinguishable at that time. So, even if we have two amoebas which share stages before the fission point t^* , there still is at any point before the fission point t^* just *one* amoeba, which “becomes” *two* amoebas only afterwards.

5.3 Persons

In the last section we discussed quite arcane issues, which sometimes may even seem to be “off topic”. However, we had to face these issues and puzzles as they regularly confuse the discussions about persons and personal identity in the (philosophical) literature. Our aim was to disentangle questions that may be dealt with by clarifying the notion of identity in general from questions that are specifically related to the notion of person and personal identity. We can breathe more freely now, after having shown how many issues can be resolved by adopting the view of perduring objects as 4-dimensional continuants.

The present section is devoted to the question as to what conditions a continuant needs to fulfil to be recognised as a person. We require an account of these conditions in order to judge whether certain forms of intervention, by affecting particular capabilities, could deprive persons of personhood.⁵² As mentioned in the introductory section, the case that a person is

⁵⁰ After a fission point, our very language changes, as we have *two* names at our disposal then where there used to be only one. It is this “new” language that enables us to give a different account of the past.

⁵¹ Note that this is extremely plausible. For, else, to correctly count objects, we would have to know the future!

⁵² It seems obvious that the reasons *why* we have to be able to judge effects of interventions on personhood ultimately have to do with the very foundations of ethics itself. Roughly, the connection to ethics is this: a) persons usually hold the very conditions securing their own persistence *as* persons in high esteem and b) it is *persons* who discuss, decide on and are addressed by moral and legal norms.

annihilated because constitutive conditions for personhood no longer obtain needs to be distinguished from the case that a person is annihilated because of a change of personal identity. In both cases the existence of a person comes to an end, but only in the second case a new person (however deficient) comes into existence. By acknowledging the possibility of a genuine change in personal identity it is clear that the *criteria of persistence for persons* (specifying the conditions under which a person can be said to remain the “same”) will differ from the *criteria of personhood* (the conditions for applying the constitutive predicate “person”). For only given these differences we can imagine that, after having undergone an intervention, someone still may be considered to be a person even though his or her former identity has been wiped out. So, we will tackle the question of the criteria of personhood in this section, while in the next we will frame the criteria of persistence for persons.

5.3.1 “To Consider Oneself as Oneself” – A Transcendental Analysis of the Concept of a Person

We won’t live up to the conceptual challenge at hand by just enumerating certain conditions we feel someone has to fulfil to count as a person. This approach would be too *ad hoc*. Rather, we should first answer the question what the concept of a person is actually needed for, what it is supposed to *distinguish*. Having answered this we may then *derive* the different aspects which, otherwise, we would only enumerate.⁵³ The starting point for our analysis will be the one that has been constitutive for the philosophical debate on the concept of a person in modern times. The core ideas of the modern history of deliberation about persons and their identity have been laid down by John Locke in his “Essay Concerning Human Understanding” (1975 [1689]). First of all, Locke clearly distinguishes *human beings* (“man”) from *persons*, the identity of the former being founded in their living bodies (1975:II, 27, §6), while the identity of the latter is constituted by “consciousness” (ibid.:§9). Locke’s second important insight is that “person” is a “forensic term” or, as we shall prefer to say, a *moral category*, “appropriating actions and their merit” (ibid.:§26).⁵⁴ Basically, a person is a being that can be held *responsible* for its actions. Therefore, it is not surprising that the concept of a person will always be encountered in contexts of rights, duties and law.

⁵³ Also, a mere enumeration would always leave open the question of its completeness. On the other hand, while every list of conceptually derived aspects too may be open to further enlargement, that would not be a defect: for there would be clear criteria for whether an aspect has to be included in the list or not.

⁵⁴ Locke was not the first to frame this idea, but he is still a particularly important historical warrantor, since much of the contemporary thinking on persons traces back to the enormous influence of his work. For historical surveys on the person concept see e. g. Noonan (2003) and Sturma (1997).

It follows from Locke's distinction that there can be persons that are not humans and humans that are not persons. As we already pointed out in the introduction to this chapter, the latter does, of course, not at all mean that such humans – like babies – don't have *rights*. On the contrary: as "person" is tied to "responsibility", babies will just have no *duties*. In this context it should be of interest to note that "human", too, is in fact not a purely descriptive category (like "homo sapiens sapiens"). However, it rather conveys *rights* than duties and so we e.g. speak of "human rights" and not of "person rights". Think of the concept of a "legal person" applied to abstract entities like companies: the main reason for applying the concept "person" to them is that thereby they can be held responsible e.g. by being *sued*. Babies, on the other hand, can neither be sued nor prosecuted or convicted.

From the *form* of the concept of responsibility alone, we can already derive some interesting aspects of what being a person implies: On the one hand, *one* is responsible *for* one's actions, but, on the other, all this would not make any sense if this being responsible was not a responsibility one has *to* someone. So the relation of responsibility is three-place: *x* is responsible *for H* to *y*.⁵⁵ The question, of course, is what goes into the place of *y*. There may be cultures in which certain people owe responsibility only to particular groups or even just to themselves ("god-kings"), but in our modern societies, which have their roots in the universal ideas of the Enlightenment, a person is held *legally* responsible for her deeds by the whole of the community of persons falling under the laws and jurisdiction in question (usually institutionalised as a state), and held *morally* responsible for her deeds by the (ideal) community of *all other* persons. However, the question as to what degree responsibility is universalisable has no bearing on the simple fact that for someone to be held responsible for their deeds they at least must fulfil the cognitive *prerequisites* for acquiring the concept of responsibility (if we do not want to postulate that persons must necessarily *have* that concept already). These prerequisites are that one must be able to *regard oneself as oneself* and so to *attribute one's actions to oneself*. That is, a person will not only have *consciousness* (certain mental states and capabilities which are presupposed by the sheer ability to act and which we will consider in more detail later), but *self-consciousness* (having a concept of oneself as the one *having* these mental states and *doing* those actions). That is what Immanuel Kant meant with his notion of the "transcendental unity of apperception", the "I think", which "must be able to accompany all my representations" ("Kritik der reinen Vernunft", 1904/11 [AA III]:B 131).

⁵⁵ Even those who would be inclined to think that persons are responsible for their actions *simpliciter*, usually *do* have some surrogate for the third argument place in the relation of responsibility: "responsibility simpliciter" is responsibility to "God" or to "Nature" etc.

However, one cannot have a concept of oneself if one does not also have a concept of other things *besides* oneself.⁵⁶ A real “solipsistic” consciousness would not even have “I-thoughts” (“I am alone”), but rather – if at all – just “there-thoughts” (“There is hunger now”).⁵⁷ Being able to make a distinction between oneself and other things in turn does presuppose distinguishing between *how things are* and *how they appear* – the objective “facts” and the subjective “experiences” (“perceptions”) which represent them in veridical or non-veridical fashion. If a being could not make this distinction, it would not have (nor have the need for) a concept of “experience” at all. And if it had no use for the concept of experience, it *a fortiori* would have no need to attribute experiences *to* anything – least of all to *itself*. One might think that a being might come to distinguish itself from other things by merely recognising the “unique” position of what is usually called “one’s own body”. However, first of all, having the concept of a *body* does already presuppose a host of other conditions, among them to have at one’s disposal the concepts of spatial objects and object-permanence, and, in fact, everything that is also presupposed by having the concepts of the “objective” and the “subjective” (as we will see soon). Secondly, the *uniqueness* of one’s body is precisely a uniqueness with respect to the systematic relation it bears to *experiences*. So, there is no way to get to the concept of “oneself” by “circumventing” the concept of experience.

⁵⁶ For readers familiar with Strawson’s “Individuals” (1959) and “The Bounds of Sense” (1966) it should be obvious how indebted we are to him in this and in what follows. That our course of argument on the whole still differs from his is largely due to Strawson not starting his presuppositional analysis at the same point as we do. That is, he does not consider the question regarding the conditions of the possibility of someone being able to acquire the concept of responsibility, but rather starts out with analysing the conditions of the possibility of identifying particulars (in “Individuals”) or the conditions of the possibility of experiential knowledge (in “The Bounds of Sense”). Unfortunately, he is therefore led to systematically neglecting Locke’s fundamental insight that personhood is primarily a moral category. In today’s discussions about the concept of a person, Strawson’s contribution is usually reduced to the thesis that persons are particulars to which M-predicates (physical predicates) as well as P-predicates (person-specific mental predicates) apply. As this – at least at first glance – has the look and feel of a somewhat bloated triviality, to readers of secondary literature it may often seem that Strawson in fact hasn’t contributed much of value to the topic. This impression is only compounded by Strawson insisting to say that the concept of a person is “logically primitive” or “unanalysable” – statements which unfortunately are prone to be completely misunderstood as entailing that the concept of a person cannot be subjected to philosophical analysis. However, what Strawson actually means is that persons are not composites of two different kinds of things: body and soul. And, rather than by the dull thesis that persons exemplify both M- and P-predicates, the richness of Strawson’s concept of a person is much better expounded by saying that persons are the subjects of experiential knowledge (in the Kantian sense) who *as such* necessarily have to be embodied. See Strawson (1966:II.1.).

⁵⁷ See e. g. Strawson’s refutation of the “no-subject doctrine of the self” (1959:Ch. 3).

Having a concept of (subjective) experiences *of* (objective) facts is the very same as having a concept of experiential *knowledge*, which is the starting point of Kant's epistemological endeavours. Of course, it would not make sense to ask about the conditions of a being having the concept of experiential knowledge if there were no such knowledge to be had. Remember that, when we analyse the concept of a person, it is ultimately *us* (i.e. the community of all those who are able to interact with each other and engage in mutual commitments, including but not limited to the community of the authors and potential readers of this paper) we are talking about. And we cannot but presuppose that the conditions of our having the *concept* of experiential knowledge are ultimately the same conditions which – in obtaining – secure that experiential knowledge is itself possible. The difference between the question of the conditions of the possibility of a person having a certain concept and the question of the conditions of the possibility of that concept having applications is just this: the answers to the former will tell us something about cognitive features of *persons*. The latter will tell us something about the general features (*synthetic a priori*) of the *world*. And that the latter fits the former so tightly is due to the “world” being the world *of* persons, that is, the world as it is experienced/constituted *by* persons. Thus, we can adopt Kant's transcendental approach by asking “How can one have experiential knowledge?” To lead to knowledge at all, the content of one's experiences must be expressible in statements, and thus experiences must be subsumable under concepts. For this to be possible, the candidate possessor of empirical knowledge must have certain *recognitional* capabilities (a subclass of memory abilities) pertaining to certain recurring features of experience. Otherwise (i.e. if there were no such recurring features or the being would lack these recognitional capabilities), no concepts could ever be formed and no experiential knowledge achieved. The indispensability of recognitional capabilities and recurrent features of experience may be expressed by saying that experiences must feature a certain *regularity* or *unity*.

Note, that while the recognitional capabilities do presuppose the existence of permanent (in the sense of recurring) features (“Q again”), they do *not* yet secure the permanence of individual *objects* of experience (“a again”), whose indispensability we promised to establish in Section 2 (see footnote 32). This task we will resume now: That experience requires as a condition of its possibility some recognitional capabilities already trivially implies that experiences are arranged in a time-order (actually, they *constitute* this time-order, but we can neglect this subtlety in the present context). What's more, for objective facts to be distinguishable from the subjective experiences one has of them, they must constitute a time-order *of their own*. This presupposes that the subjective order of experiences is not always *necessary* – in the sense that one could, at least in some cases, possibly experience things in reverse order (for example, one may first look at the desk and then at the chair, or first at the chair and then at the desk). This, in turn, presupposes

two things: that different possible objects of experience can exist *simultaneously for at least a while*, and that this can be the case *independent* of one actually experiencing them or not. The first condition is what affords an at least relative *permanence of objects* of experience over time (thereby finally proving the view to be incoherent that concrete objects may exist at *single* specific points in space-time only, and that permanence through change is a concept that can be dispensed with). The second condition entails that the objects of experience must exist within more than just a *time* frame. The “more” can be tentatively identified with space, even though the argument itself does not show that the extended framework must have exactly four dimensions or must in other respects be anything like actual space-time. However, it is the relative object permanence again which is *also* the condition for the constitution of the spatial frame and the location of “places” within it – think of how we use relatively permanent geographical features and relations on Earth in this way, or think of our reference to “fixed stars” for defining spatial locations within universe itself.

The possibility of time orders (subjective or objective) trivially presupposes that experiences differ, that is, that there are changes. However – and this is by no means trivial – the changes occurring have to be compatible with what we just established to be a condition of the possibility of experiential knowledge, namely (at least relative) object permanence.⁵⁸ And while this will not give us the strict *principle of causality* (“every event has a cause”) it will establish something like a *principle of sufficient causality*, that is, the connection of changes in experience calls for an extent of regularity which does suffice to recognise objects after a time of (unobserved) change as being still “the same”.⁵⁹

Let us sum up what we got thus far: We analysed the concept of a person in terms of a being responsible for its actions. We maintained that, even if we do not expect such a being to have at its disposal the very concept of respon-

⁵⁸ In his famous book “A Treatise on Human Nature” (1739/40), David Hume maintained that the application of the concept of identity must necessarily be confined to unchanging objects. He surmises that this is actually part of what motivated the (mistaken) idea of *substances*: If an object perdures, then through all changes of *attributes* the *substance* in which the attributes “inhere” (and which, in essence, *is* the object) must stay unchanged. In Hume’s view, however, there are no substances and, therefore, there is no object permanence (cf. Hume 2000: especially I, 1.6., 4.2., 4.6.). It is important to realise that Hume is not correct in maintaining that the concept of object permanence does presuppose the concept of underlying unchanging substances – this would be the case only if he were right that the notion of identity cannot be used in application to changing objects in the first place. But, as we saw in Section 2, identity *can* be reconciled with change – and thus a consistent notion of object permanence established – by drawing on the concept of continuants as objects having temporal parts.

⁵⁹ It is clear that recognition despite change would not work if changes always occurred in an irregular and unforeseeable fashion (imagine a cup suddenly transforming into a guitar while no one is looking).

sibility itself, it at least needs to fulfil the conceptual prerequisites necessary for acquiring that concept. We found that the decisive prerequisite is to have a concept of *oneself* in order to be able to attribute one's actions to oneself. This again presupposes a being to have the concept of (subjective) experiences of (objective) facts, that is, the concept of experiential *knowledge* – the applicability of which in turn presupposes relative object permanence and sufficient causality within the framework of an objectively constituted space-time.

However, *all this*, while indeed expounding necessary conditions of personhood, is still not sufficient for establishing that a being that meets these conditions will really have a concept of *itself*. The reason is that the distinction between *how things are* and *how they appear to be* does not by itself explain why experiences are attributed to a subject having them.⁶⁰ The aspect still missing from the picture is that having the concept of *oneself*, of someone who is a subject of experiences of objective facts, presupposes in addition the concept of “others” – that is, “others like me”.

This becomes evident when reflecting on the fact that the conceptual realm presupposes language and meaning. For language and meaning are constituted by rules that require a social setting for their working. Here we draw on Wittgenstein's famous argument against the possibility of “private languages” (2001 [1953]:§243ff., especially §258). Roughly speaking, the argument runs as follows: The meaning of any expression is constituted by its usage as determined by rules. Speaking is actually acting according to rules. Rules, in general, require intersubjective (“public”) criteria in order to decide on their being followed or violated. For it is meaningless to say of a person that he or she is observing some rule R as long as the only criterion for the person being *correct* in her application of R is that the person herself *thinks* she is. In fact, saying something in this vein merely would serve to disguise that there is no criterion after all and, consequently, no rule. However, if there was a private language, then by definition the “private” objects of its application (e.g. sense-data only accessible to me etc.) would force situations of just that kind. In a private language any opinion would be “true” as long as its alleged speaker would take it to be true. Whenever someone *felt* that some “sensation” was the same as some other sensation, there would be no grounds for opposition. This kind of absolute epistemic certainty is but an apparent advantage of private languages. It actually shows that for putative languages of that kind there is no way of telling “right” from “wrong”, thus rendering them meaningless.

⁶⁰ This is shown by Strawson (1959:Ch. 3). He furthermore argues that the unique position of what is later to be called one's “own body” still does not warrant the attribution of experiences to anything. Husserl would have said that a phenomenological investigation of the issue reveals that experiences have an *intentional* structure and that this structure includes a “subject pole” as well as an “object pole”. That is surely true, but we want to know *why* this *has* to be so.

So, to make use of language and to confer meaning is not a private affair; it rather presupposes being part of a *community*. Consequently, being a person ultimately presupposes the acknowledgement of other persons besides oneself. This has an immediate consequence which will again deserve our attention later on in Section 5.4: for it entails that persons must be able to recognise (identify) each other. Hence, as subjects of experiences, persons do not only experience the “objective world” from their different “subjective perspectives”, but – even more important – they are the *objects* of the experience of *other* persons with whom, by communication, they intersubjectively constitute a common objective world.⁶¹ That is, persons mutually constitute each other as inhabitants of a common objective world, and as such they must be locatable within the objective space-time framework too, running their course as long as they persist.

Starting with the rather parsimonious supposition that a person must be able to at least *acquire* the concept of responsibility, we came a long way in deriving a number of further conditions a being has to fulfil for being recognised as a person: Above all, it must have a concept of *itself*, which in turn demands cognitive capacities to such an extent that it can be said to be a subject of experiential knowledge, a participant in a community of language users, and as such locatable within the common objective world the community intersubjectively constitutes. It may come as a surprise to find that, in the end, this result comes very close to Locke’s straightforward definition of a person as “a thinking intelligent Being, that has reason and reflection, and can consider it self as it self, the same thinking thing in different times and places” (Locke 1975:II, 27, §9).

⁶¹ It becomes obvious now that the concept of objectivity featured throughout this text does not entail realism, but rather is neutral regarding this issue. While some authors (most prominently, the “early” Husserl of the “*Logische Untersuchungen*” [1900/1901]) indeed endorse a realistic understanding of objects of experience, they can be just as well understood as being constituted in their “objective being” as an outcome of discourse among subjects of differing experiences, i.e. persons. Objectivity in this sense is properly understood as *intersubjectivity*. And, though in the opinion of Kant and Husserl the feat is to be accomplished via general *conditions* regarding the subjects of experience rather than by discourse among them, intersubjectivity is also the sense in which both (Husserl since the “*Ideas*” [1913]) conceived of the “objectivity” of intentional objects of experience. While this, at first glance, may seem to be at odds with Kant’s famous distinction between “things in themselves” and “things as they appear”, it is important to realise that in writing on objectivity as secured by the categories of the understanding (in the “*Transzendente Analytik*”), Kant wasn’t concerned at all with arguing for the existence of inexperiencable “things in themselves” outside of the realm of the “*Anschauung*”, but for the fact of objectivity *within* that empirical realm itself. (It is remarkable that the only – fairly weak – argument regarding the additional existence of “things in themselves” *outside* the realm of experience is given in the preface to the 2nd edition of the “*Critique of Pure Reason*”. It bluntly states that we have to assume the existence of things in themselves to avoid the “contradiction” of there being “appearances” without anything that is appearing.)

5.3.2 "Being Capable of Having One's Actions Imputed to Oneself" – Further Cognitive, Motivational and Emotive Requirements of Personhood

In the following, we will try to dig a bit deeper still. That is, we will mainly talk *in more detail* about the cognitive capacities which we established above as conditions of personhood, and we will broaden the scope by considering motivational and emotive capacities as well. To this end, we once more begin with Kant, who confirms the forensic account of personhood by stating: "Person ist dasjenige Subjekt, dessen Handlungen einer Zurechnung fähig sind."⁶² Kant stipulates *freedom of the will* or *autonomy* as a decisive prerequisite for the kind of imputability that he is considering to be constitutive for personhood. Autonomy stands for the freedom of rational agents to subject only to laws they adopt themselves by exercising reason. And reason, in turn, is viewed by him as located *outside* the causally determined realm of experience, even if it is – indirectly – operating on it. At this point it may seem as if we could not avoid entering the gargantuan field of debate about free will and its relation to the rival positions of (causal) determinism and anti-determinism. But fortunately we can. Most extant determinists are so-called *compatibilists*. Still, compatibilist determinists and incompatibilist libertarians *agree* that there is free will and responsibility and that the former is a necessary condition of the latter.⁶³ This opens up the possibility of formulating a position which is *neutral* with respect to the metaphysical issue of whether determinism is true or not.⁶⁴ In the following, we shall examine in neutral

⁶² "Metaphysik der Sitten" (Kant 1907/14 [AA VI]:223) – "A person is a subject who is capable of having his actions imputed to him" (translation by W. Hastie). – Taken at face value, this quote might seem to imply that persons are subjects of a certain kind, i.e. constituting a subclass of subjects. However, this interpretation does not fit into the general frame of Kant's philosophy. The way he uses the two terms rather lends itself to the interpretation that all subjects (at least insofar as being a subject is meant to entail being capable of arriving at experiential knowledge) are persons. According to our own analysis, it furthermore seems that "person" and "subject" are not only coextensional predicates, but semantically necessarily so, that is, they in fact express the same concept. Their usage differs only insofar that the expression "subject" emphasises epistemological (theoretical) aspects, while "person" indicates that attention is to be focussed on moral (practical) aspects.

⁶³ That is because compatibilism holds that free will and responsibility are *compatible* with determinism. Only a minute fraction of philosophers contend that determinism is true and *hence* there is no free will and *hence* no one bears responsibility for his or her actions. Sometimes, these philosophers then advocate radical changes in our legal and juridical system to the effect that people should not be sentenced and punished according to "obsolete" categories like guilt, retribution or penance, but rather with respect to the chances of their resocialisation (i.e. the probability of them abstaining from committing crimes in the future). Of course, one cannot but wonder what performative sense such appeals to reform do make against the background that, if determinism is true, what is bound to happen will happen anyway.

⁶⁴ In Hartmann (2000/2005) it is argued that determinism is in fact false.

terms what kind of freedom is required for persons so that we may hold them responsible for their deeds.

Our whole practice of blaming or praising each other (morally or legally) for what we did is based on the conviction that – at least under normal circumstances – we could have done otherwise, had we just chosen to do so. By employing this practice we treat each other as persons, that is as beings who are free in a certain respect that can be addressed by the following condition:

Ia) Persons need to be able to *refrain* from actions.

Complementary to this condition of personhood is a second one, namely:

Ib) Persons need to be able to *act out their decisions* to act.

Both these conditions are already implicit in the concept of *agency* itself. That is, beings qualify as agents by virtue of these abilities, and the range of things they do, by which they can exercise these abilities, is the range of their actions.⁶⁵ Being an agent, as compared to a mere stimulus-drive-response automaton, means to not be determined exhaustively by external forces and internal drives. To illustrate this point further, it is helpful to introduce a terminological distinction between behaviour and action (which is not followed very strictly in ordinary language): Responses of persons can be subdivided into two classes. We suggest to call those responses that *merely happen to us* – like coughing, stumbling or crying – *behaviour*. Every now and then we may successfully *suppress* this kind of responses, but only for *actions* it is apt to say that we can *refrain* from exhibiting them. This intuitively plausible distinction is strengthened by the different ways of accounting for these two classes of responses in science. While behaviour – at least in principle – is predictable by causal laws that are established by physiology, ethology, and behavioural psychology, this is not the case for actions (Hartmann 1998:42–45).⁶⁶ That does not mean that actions are events that cannot be the objects of systematically correct predictions. Certainly, one can establish “predictors” for actions as well, insofar as the likelihood for persons to respond with a certain type of action in a situation of a given type usually depends on a number of internal and external factors – especially their system of goals and beliefs. Yet, these factors do not *cause* actions in a strict terminological sense

⁶⁵ It follows that not everything an agent “does” is an action. E.g. if you are reading this chapter, you must be an agent. But if you have to sneeze or cough while doing so, these are not “actions” of yours.

⁶⁶ The modern habit of classifying actions as a special (“complex”) class of (“goal-directed”) behaviour has its forgotten roots in behaviourist psychology and thus is reductionistic and deterministic in spirit. However, it is important to see that the classification of a response as an action goes along with the application of a whole range of other termini (“purpose”, “intention”, “belief” etc.), so that actions are embedded in an altogether different theoretical context compared to behaviour (which is accounted for by using termini such as “reinforcer”, “drive”, “stimulus”, “conditioning” etc.).

(Hartmann 2000:82). Furthermore, there are factors like drives and emotions, which certainly will influence the decisions of agents (for instance, agents who have the choice between finishing some work and accompanying a friend to a restaurant will be more inclined to opt for the latter if they are hungry). But as long as we are considering an event as an action, we take for granted that the agent could have refrained from it, no matter how strong the motivating drives or emotional incentives involved. The same goes for the counterpart, the “acting out” of a decision against opposing drives or unpleasant emotions. This is the reason why we blame agents for their actions or their omissions, but not for their mere behavioural responses.

Compared to Kant’s categorical concept of autonomy, the kind of freedom that is postulated by (Ia) and (Ib) allows for degrees. Depending on the extent to which persons actualise the capabilities mentioned in these conditions they can be said to possess a stronger or weaker will.⁶⁷ Personhood, on the other hand, does not come in degrees. At a given time, a being either is or is not a person. However, there is no need to specify a certain degree to which an agent needs to develop his or her will power in order to be recognised as a person. If someone is at all able to act and capable of taking responsibility for his or her actions, then he or she is, in this sense, a *moral agent*, i.e. a person. Given these abilities, people are to be recognised as persons regardless of whether they actually act morally, whether they engage in moral deliberations on what ought or ought not to be done, or whether they exercise strong self-control over their inclinations and desires. At this point it might seem about as difficult to judge whether someone has the abilities (Ia) and (Ib) as it was, in the first place, to judge whether someone is a person. However, our account will become more informative once we proceed with our analysis by asking what cognitive and emotive characteristics a being needs to be endowed with in order to develop the abilities to act or refrain from action on decision.

As a first step to further our analysis of personhood we state some rather basic cognitive abilities that, quite uncontroversially, are necessary prerequisites for (Ia) and (Ib):

IIa) Persons need to be endowed with discriminative abilities.

In talking about discriminative abilities we refer to *perception* as well as to *recognition*. The reason why we subsume these abilities under this heading is that perception and recognition are *ascribed* to beings on the basis of the discriminations they make in their responses (comprising actions as well as behaviour). To cut a long story short, if some being responds (or is disposed

⁶⁷ Since ancient times persons were considered virtuous if they developed these abilities to a high degree. The virtues associated with (Ia) were called temperance or self-control (German: “Besonnenheit”), the ones associated with (Ib) decisiveness or fortitude (German: “Tatkraft”). Taken together, they sometimes are quite aptly named “executive virtues”.

to respond) discriminatively to a presently given difference, then we may say that it *perceives* that difference. And if we can *account* for a being's responses towards a presently given difference only by assuming that it has been exposed to it before, then we may draw the conclusion that the being *recognises* that difference.⁶⁸

In fact, most if not all perceptions involve recognition, and recognition is in turn the most basic function of the faculty of memory, be it in the form of recognition of *types* of situation ("F again") or of objects ("a again")⁶⁹. If a being does not have recognitional capacities, it is not a person, not so much because there is a direct connection between recognitional capacities and executive virtues (though there is), but rather because a being that does not perceive/recognise much of what is going on around it will not be able to exercise very subtle forms of behaviour, let alone action. Still, obviously perception and recognition do not exhaust what is required for personhood. Taking responsibility for one's deeds presupposes another form of memory, the ability to remember *what one did*. For example, a decision one made is an event in one's life. If one cannot remember such events, how is one *ever* to act out one's decisions? Occasionally, we blame persons specifically for their forgetfulness, for instance when someone forgets having given a promise. But this is because we assume that the person would have remembered if she had only exercised some care. On the other hand, beings which cannot ever remember what they have done in the past will not be treated as responsible for their responses at all. Not only will they not be able to commit themselves to anything by promise or assent (because they will forget at once that they did); they will obviously be unable to lead any form of autonomous life. So, persons need to be able to recall their actions and, more generally, past experiences, that is they need to have *episodic memories*:

I Ib) Persons need to be endowed with episodic memory.

Closely related to memory is another requisite for personhood:

I Ic) Persons need to be endowed with learning abilities.

Quite often it is only a matter of point of view whether we refer to a being's memory abilities or its learning abilities in order to explain its responses. If we want to emphasise the process of *acquisition* of a behav-

⁶⁸ See Hartmann (1998:II.2.1) and Galert (2005:Ch. 2) for a more elaborate account. One important advantage of that account is that it provides a fairly clear understanding of how to ascribe these psychic functions to a being even in situations where (because of a lack of communicative abilities on the part of the respective being) it is not possible to simply ask it what or whether it is perceiving or recognising.

⁶⁹ It is important to see that (despite the examples in brackets) these basic functions of recognition do not yet presuppose the use of language. That is, "recognition" in this usage does convey the same meaning as the German "wiedererkennen", but not the same as "erkennen" in the sense which results in "Erkenntnis" ("knowledge"). The latter has to be of propositional structure and is therefore bound to language.

itorial modification (in the widest sense), we talk about learning. If, on the other hand, we want to emphasise the *retention* and *retrieval* of a behavioural modification, we talk about memory. A recourse to basic forms of learning is already involved at least in the *scientific* ascription of discriminative abilities as expounded above under IIa – for it is primarily in the context of learning (conditioning) experiments that we can acquire data that semantically warrant the ascription of specific abilities of perception and recognition.⁷⁰

Neither discriminative abilities, nor simple forms of learning and memory require possession of language on the part of the discriminating/learning/remembering being. Even episodic memories need not necessarily assume a linguistic form. On the contrary, we would find it rather strange if we could “tell” what we did, but would lack any accompanying *re-presentations* in the visual, tactual, auditory or olfactory modes of imagination. However, there are nevertheless independent reasons for stating that:

II d) Persons need to be endowed with language abilities.

To hold a being responsible for what it does we need to assume some understanding of what it means to be responsible for one’s deeds on the part of that being. One important aspect of being “capable of having one’s actions imputed on oneself” is to be able to get engaged in the kind of interaction that is constitutive of *interpersonal relationships*. These relationships comprise, *inter alia*, the assertion and justification of claims against each other, the making of commitments, the granting of rights, the exchange of reasons for one’s actions, the justification or questioning of actions with respect to rules, or the justification and questioning of those rules themselves. All these features of interpersonal exchange require the medium of a shared language,⁷¹ however keeping in mind that the possession of language abilities is not to be equated with the ability to speak. The different systems of sign language for deaf mutes, for instance, basically fit the same purposes as do spoken and written forms of language.

Finally, language abilities are indispensable for persons as they are the decisive prerequisite for a final cognitive ability that we consider to be a necessary condition for personhood:

II e) Persons need to be endowed with deliberative abilities.

⁷⁰ For a more general account of different forms of learning see Hartmann (1998:II.1.1.2). The relationship between the psychology of learning and the psychology of memory is dealt with *ibid.*:II.2.5.

⁷¹ For the purposes at hand we do not need a clear-cut distinction between mere *communicative* abilities and *language* abilities in the strict sense. For a tentative account of some distinctive features of language see Hartmann (1998:167–168). However, it is clear from the examples given in the text that the communicative abilities of (at the very least) most animals lack the complexity required for typical interpersonal interaction.

No doubt, there are theoreticians who stand up for the possibility of thought without language (see e. g. Weiskrantz 1988). Once again, we cannot delve here into the reasons why we consider this position to be ill-founded, but we set them forth elsewhere (see Hartmann 1998:172–176). The need for deliberative abilities is already implied by what we just said concerning interpersonal relationships. So, we only would like to add that it is mostly *by deliberation* that persons arrive at decisions to act in spite of internal or external hindrances, or at decisions to refrain from acting in spite of certain immediate satisfactions or benefits that an action might bring about.

Considering the cognitive requirements for personhood we deduced above, it is obvious that animals will hardly qualify as persons and that infants are not *yet* persons. Animals as well as infants may be considered to have “personalities” in the weak sense that they can have individual character traits, that is certain relatively stable bundles of dispositions for behaviour of a certain kind (especially traits that are traditionally regarded as constituting the “temper”, like aggressiveness, irritability, anxiousness, curiosity etc. – see Section 5.4.3 for more on this). But we will limit our use of the term “personality” to the character of persons. Our account of the cognitive criteria for personhood also explains why, in the introductory section of this chapter, we came to the conclusion that human beings in persistent vegetative state can no longer be regarded as persons. For these people as well as severely mentally challenged people do not fulfil most of the conditions we found to be constitutive for personhood. However, there is no doubt that, according to our criteria, most people who are considered to suffer from some mental disability will have to be regarded as persons as long as they are able to get engaged in social interactions to such an extent that they can grasp the meaning of responsibilities and mutual obligations. After all, a requirement like the demand of language abilities is not to be understood in such a way that a person would need to be able to comply with the highest standards of linguistic perfection. Finally, we also would like to reinforce the point we made above that by denying a human being the status of a person he or she is deprived of all *duties* and *responsibilities*, but not of his or her moral and juridical *rights*.⁷² A common way of stating this is to say that the range of *moral objects* is broader than the range of *moral subjects*. Furthermore, it should be noted that by distinguishing between the moral category “person” and the biological category “human” the possibility is left open for there to be persons that are not humans as well. It is only a contingent fact that so far we never came across any nonhuman beings that would fulfil the cognitive criteria for personhood.

With our account of the cognitive prerequisites for personhood completed, we will ask what else might be required of persons besides these cog-

⁷² Except, of course, such rights whose execution entails one’s taking over certain duties – for example the right to negotiate and submit to a contract.

nitive abilities. We already stated that being a person is by no means a private affair, and now we would like to add that it is no ghostly affair either:

III) Persons need to be embodied.

This provision in fact follows from what we established already in this chapter, namely that persons must be locatable within the frame of objective space-time through which they run their course as long as they persist. The embodiment condition might seem trivial to some, but facing the long standing debate on persons as immaterial souls it certainly is not. However, given that persons are moral agents in the sense that they are capable of having their actions imputed on them, embodiment is obviously indispensable for being a person: Even if we grant, for the moment, that immaterial beings could effectuate changes in the material world, this would at once confront us with insurmountable problems in the context of actually *ascribing* such changes to mere spirits. If you think of a typical ghost story featuring some invisible poltergeist tossing down books from the shelves and so on, how are we to hold a *particular* poltergeist responsible for these nasty deeds? After all, it could be a dozen of them! By adding some recognisable ghostly voice coming out of nowhere and commenting on the deeds, the scenario certainly gets more “personish” (and scary), but it nevertheless won’t be a candidate for serious consideration, because we would still lack the means for establishing a reliable connection between the invisible source of the voice and the intangible originator of the “actions”.

The requirement of a body does, however, *not* presuppose even a remote resemblance to the human body. The body of a person may be of any matter, structure, shape and size as long as it is endowed with sense organs (not necessarily ours) and executive organs that allow actions in general and linguistic interactions *with us* (i.e. the community of human persons) in particular. We need not even assume the existence of a brain, although – at present – we don’t know of any structure other than a brain that could warrant implementation of the functional requirements we postulated before.⁷³ We emphasised the indeterminacy of condition (III) in order to counterbalance any impression of “species chauvinism” that our account of the cognitive prerequisites for personhood might have created. On the other hand, since this study deals with interventions in human beings, the question of which degrees of freedom there are with respect to the way persons can be embodied is not of great relevance for the purposes at hand.

We will now inquire if there are – beyond the cognitive prerequisites and the embodiment condition – motivational and emotional prerequisites as well. Let’s first consider whether a person necessarily needs to have certain

⁷³ But maybe we can at least *conceive* of possible designs for intelligent “no-brainers”. Think, for instance, of a being with a web of ganglia (or just ganglia-like structures) evenly pervading its body.

motifs. Now, motifs come in many different varieties, so that a satisfactory answer to this question would require us to introduce a number of concepts. As this would go beyond the scope of this paper, we can only give some hints as to what we take to be an appropriate answer. First, being an agent, a person will necessarily entertain certain *purposes*. This is a conceptual need, for once we acknowledge a being's responses as actions, we describe them as being motivated by purposes (and, furthermore, as being guided by beliefs, see Hartmann 1998:237). By "purposes" we simply mean the states of affairs that actions are meant to bring about.⁷⁴ Accordingly, as agents, persons necessarily are motivated by purposes:

IVa) Persons need to entertain purposes.

With respect to the contingent⁷⁵ fact that all known persons are living beings, we can state some more motivational necessities: Living beings exhibit a number of responses which repeatedly bring about certain states of affairs that are of importance for the maintenance of their life or for the maintenance of their species. These states of affairs can be said to constitute a living being's *natural needs* (Hartmann 1998:50) and the related responses (if effective) can be called *satisfactions* of natural needs. Against this background, we now can say that having the disposition to satisfy natural needs is a constituent of a living being's continuing existence. Albeit not endowed with the same kind of conceptual necessity as (IVa) and the preceding conditions, we can note:

IVb) As *living beings* persons need to have the disposition to satisfy their natural needs.

Motivational aspects are closely linked with *emotional* ones and emotions are as varied as are motifs. Sensations constitute one (primitive) type of emotions that is already involved in learning. One of the basic forms of learning, operant conditioning, requires reinforcement and inhibition – "reward" and "punishment" in ordinary parlance. More generally speaking, a living being would lack the motivation to learn anything if it did not prefer certain situations over others, if it did not seek certain situations while avoiding others. We would like to call such sensations accompanying the perception of situations of reinforcement *sensations of like*. Correspondingly,

⁷⁴ Cf. Hartmann (1998:66). – By way of further analysis one can come to distinguish "interests" and "goals" as different kinds of purposes (ibid.:67). Furthermore, in order to account for actions that failed to bring about their purposes the terminus "intention" is required (ibid.:71–73). However, for the present context we do not need to introduce these distinctions properly.

⁷⁵ However, if we call to mind that any embodiment of psychic functions will require transformation of energy as well as a certain sort of "maintenance", the following considerations would, for the most part, hold for *any* persons, be they biological organisms, or systems which, for one reason or another, could not terminologically be addressed as "life forms".

sensations that a living being has while perceiving situations of inhibition we call *sensations of dislike*.⁷⁶ It is important to note that sensations, in our understanding, do not come along as a mere corollary to perceptions. Rather, the need for the term “sensation” is already given with the distinction between an “objective world” (intersubjectively constituted – see above) and the “subjective experience” of it. Perception in an actual sense only takes place when what a living being’s senses convey to it pass as a veridical “representation” of the world – one cannot perceive what is not there. However, as we all know, cases of illusion and other sorts of sensory misrepresentation abound. For the affected being these misrepresentations are often not easily recognised as such. To that being it seems as if something was the case which *in fact* is not.⁷⁷ We account for situations of this kind by saying that the being has the *same sensations* that it would have if what appeared to it was actually the case. In other words, instances of sensory illusion and corresponding instances of perception are indistinguishable sensation-wise (Hartmann 1998:116–119). Hence the requirement that persons need to have sensations already follows from (IIa), the requirement of discriminative abilities in the particular form of sensory perception. Regarding motivation, the sensations of like and dislike are particularly noteworthy as they represent a necessary prerequisite for (IIc), the ability to learn. Therefore:

IVc) Persons need to have sensations of like and dislike.⁷⁸

It seems questionable to qualify any further particular types of sensation as being necessary for personhood. Remember that, when discussing the requirement of discriminative abilities, we did not postulate that persons need to be endowed with any *particular* mode of perception. A type of sen-

⁷⁶ See Hartmann 1998:215.

⁷⁷ In order to distinguish facts from mere sensory appearance, intersubjective agreement regarding the appearance frequently won’t help. For if all those who negotiate the facts are affected by the same illusion by virtue of them sharing the same sensory apparatus, then an illusion will not be recognised as such. To get a grip on the facts it takes reference to *transsubjective* standards, e.g. the overriding status of other sensory modes, or of measurements. Only with respect to such overriding standards can intersubjectively reproducible illusions be regarded as illusions in the first place.

⁷⁸ Although our way of introducing the concept of sensation should probably have made this clear enough already, we would like to state explicitly that by (IVc) we do not identify sensations with what philosophers have come to dub “qualia” and, hence, (IVc) does not at all state or imply that persons need to be endowed with such things. Indeed we would maintain that the concept of qualia as it is discussed within contemporary philosophy is devoid of meaning. For a thorough critique see Hartmann (1998:121–123). Whereas behaviour of any complexity is alleged to be conceptually compatible with the total absence of qualia (or their varying widely between beings), and hence one can (allegedly) “never know” about the qualia of a being without actually *being* that being, the simple fact that a living being seeks/avoids certain situations semantically suffices to say that it has the according sensations of like/dislike.

sation would classify only if it could be shown that its loss would imply the negation of one of the conditions Ia–IVc. To test this, let's examine one *prima facie*-candidate: Given the presupposition of self-consciousness, the need for discriminative abilities, and condition (III), that persons need to be embodied, it may seem quite conclusive that persons need to possess proprioceptive abilities. Else they would perceive their own body, especially the orientation of their body parts to one another, only through the “outer” senses, and how could they then exercise the sort of cybernetic control needed for the execution of even the most basic actions? However, this reasoning is challenged by some (very rare) cases of people who suffered a complete, yet selective, loss of the sense for their own bodies due to a viral infection. Certainly, these people are severely disabled since vision is their only source of feedback for controlling movements of their bodies. Only by arduous exercise can they reacquire a limited range of bodily actions. However, the published reports on such cases leave no doubt that these patients, in spite of their disabled condition, still show all the marks of personhood we established above.⁷⁹ Maybe another candidate to consider would be the capability to feel *pain*, which is a particularly salient type of sensation of dislike.⁸⁰ Still, cases of congenital insensitivity to pain (see e.g. Sternbach 1968:95) clearly show that sufferers from this sensory deficit, as unfortunate as they are, nevertheless satisfy all criteria of personhood. That neither the possession of bodily sensations in general, nor of pain sensations in particular passes the test for being a prerequisite of personhood seems to us sufficient evidence for the assumption that in fact no particular mode of sensation is thus required.

Also, the realm of *sentiments* does not offer any firm grounds for postulating further requirements for personhood. A person that does not feel joy, sorrow, fear, anger, disgust, despair, hate, love, etc, surely will strike us as a person with a serious disorder, but will still be a person. It may be asked whether the capacity of having sensations of like and dislike together with the capacity of having purposes and linguistic abilities will not conceptually ensure that there will be at least *some* sentiments, even if there are no specific ones. As a warrant for this we could cite that a host of sentiments are constituted at least in part cognitively through the interpretation of one's sensations (e.g. accompanying the perception of one's vegetative state) in the light

⁷⁹ One such case was described by Oliver Sacks in his famous collection of neurological abnormalities “The Man Who Mistook His Wife for a Hat” (1985), under the heading “The disembodied lady”. In another case study, Jonathan Cole (1995) portrayed the heroic struggle of a man who lost his sense of proprioception at the age of 19. Both these patients *acquired* their peculiar sensory deficit. Accordingly, one could still raise the question whether a human being that is *born* without proprioceptive abilities could at all *develop* into a person.

⁸⁰ For those who are in doubt whether pain sensations are necessarily to be disliked, see Galert (2005:3.3.4.).

of situational factors (“causal attribution”), which then leads to further responses, etc. However, we are inclined to think that this does not suffice to show that persons actually *must* have sentiments thus constituted. People may lose or altogether lack the ability to make causal attributions pertinent to emotions without at the same time lacking the cognitive and motivational capacities required for personhood. Or something may be wrong with their physiology in such a way that they lack the background of bodily sensations as a basis for further interpretation in the first place.

Since we started out in our conceptual analysis by designating persons as moral agents, so-called moral sentiments – such as contempt, indignation, guilt, shame, resentment and compassion – may seem to deserve some special consideration. However, we continued by asking for the prerequisites for *acquiring* the concept of responsibility, emphasising that a person need not necessarily be virtuous or, in fact, act according to moral standards of any kind. Hence, it would be a mistake to demand of persons that they need to have moral sentiments, for such sentiments will arise only if a person actually *adopts* the concept of responsibility as her own. In the same vein, a lack of empathy does not disqualify anyone as a person, either. For instance, one symptom of autism is a more or less severe disability to “read other people’s minds” by inferring emotions from facial expressions and the like (see e.g. Baron-Cohen 1997). An impairment of that kind will severely affect one’s *de facto* responsiveness to others’ needs, but it does not as such negate any of the conditions of personhood. And though it will make it more difficult, it certainly does not preclude paying due respect to others, attending to one’s duties and so on. Therefore, provided someone suffering from autism meets the other requirements of personhood, there is no reason for denying him or her that status on account of his or her lack of empathy.⁸¹

Compared to the cognitive requirements for personhood, the demands regarding emotional and motivational traits appear to be fairly low. However, this only underscores the point we repeatedly made that the conditions for being and remaining a person do differ from the conditions for remain-

⁸¹ This might be a good place to remind the reader that the connotations of expressions like “withholding/denying a status” and “conferring/granting a status” can easily be misleading – the first sounding intrinsically like a sanction, the second intrinsically like a benefit. On the basis of such connotations, it might seem mandatory to be always and only cautious when “denying” someone person status. However, to confer person status to someone who does not satisfy the prerequisites would in truth not be a beneficial deed *at all*. For it would mean regarding them as fully responsible even though they in fact can’t help it! As an example, think of cases where parents, essentially because they mistakenly interpreted the responses of their much too young babies and toddlers as intentional actions in the full sense, believed to be fully warranted in resorting to harsh measures of “punishment” or even “retribution”. So, depending on the circumstances, withholding person status can actually mean protecting someone when conferring person status would mean doing them harm.

ing the *same* person. For considerable change in someone's emotional/motivational profile may indeed give rise to concerns regarding the persistence of his or her personal identity. This will be so, even if there is no concomitant change in the cognitive domain. These considerations already link to the subject matter of the next section, but before we finally try to clarify the criteria of persistence for persons it will be helpful to summarise, briefly, our account of personhood:

Being a person basically means to be capable in principle of entering into the sphere of mutual commitments and thus taking responsibility for one's actions. Being a person means to be a moral agent, that is, a (potential) subject of duties, in contrast to a moral object who only has rights. Holding on to this pivotal point, in a transcendental approach we inquired about the conditions of the possibility of beings of that kind. The concept of responsibility presupposes a community of beings each having a concept of themselves, that is self-consciousness, as well as a concept of "others like me". Having established the general framework allowing for the existence of persons, we then changed the perspective by putting the single person into focus. We asked what psychic conditions each and every moral agent needs to fulfil in order to be able to take responsibility for his or her actions and hold others responsible for their actions. The concept of agency in and of itself requires a certain kind of "freedom of the will" on the part of possible agents, that is they need to be able to exercise a degree of control over their responses: At least in some situations they need to be able to act on decision, but also to refrain from an action. We then proceeded by enumerating cognitive prerequisites for persons as moral agents, from the rather basic to the more exclusive ones. The basic skills – which persons share e.g. with many animals – are discriminative and learning abilities. The more exclusive ones comprise episodic memory, language and deliberative abilities. Commenting on the debate on persons as immaterial souls we established that persons need to be embodied. Finally, we considered the motivational and emotional conditions that persons need to fulfil. As agents, persons will be motivated by purposes. Furthermore, persons will be motivated to satisfy their natural needs, at least when we consider living beings. Pertaining to emotions we found only a very basic condition to be tenable, which is that persons need to have at least sensations of like and dislike. As mentioned before, we do not claim that our list of necessary conditions for personhood is in any sense complete. However, any suggested further provision will need to be considered in the light of its necessity for enabling a being to take responsibility for its actions.

To conclude this section, we would like to address some possible doubts concerning the *practical applicability* of our considerations. After all, we postulated a number of necessary conditions for personhood which all relate to abilities that allow for *degrees*. Consequently, the question arises how to decide, when actually confronted with people who are severely impaired in one or more of the relevant abilities, but without having lost them alto-

gether, whether they are still to be regarded as persons. We clearly have to admit that we will not be able to frame very explicit criteria that would allow to decide each and every borderline case of that kind, not because we wouldn't be eager and willing to do so, but rather because we are in doubt whether the problem of recognising someone's status as a person is approachable, in principle, through something like an operationalised aptitude test. However, we can offer some hints as how to handle borderline cases.

As we saw, the concept of personhood is constituted in a context of mutual recognition. If someone is not at least implicitly⁸² able to *call for recognition* as a person, the community of persons will hardly have a reason or – depending on the way you look at it – the right to convey to him or her the person status. This already settles the problem of further specifying the required language abilities for personhood. For, if someone has got the linguistic means (mind you, not necessarily means of speech!) for raising the claim to be recognised as a person, he or she already demonstrates thereby sufficient mastery of language. As language and deliberative abilities are most intimately linked, this line of argument can be applied to the latter as well: If someone has got the deliberative resources to understand what is at stake when his or her status as a person is – explicitly or implicitly – under discussion, and as long as someone is capable of standing up for their recognition as a person, they fulfil requirement (Ile) to a sufficient extent. Now, if someone actually calls for recognition as a person, it will certainly be a delicate decision to nevertheless deny them that status.⁸³ Reasons for doing so, with respect to the cognitive domain, will most likely relate to the requirement of episodic memory, for a violation of the other conditions within that group will hardly leave language abilities unaffected.

Suppose now somebody's episodic memories are utterly and irretrievably wiped out in a case of severe retrograde amnesia. Persons with that condition will no longer be the same (at least we will argue for this view in the next section), but as long as their other cognitive abilities remain unaffected, and provided they are able to store new episodic memories, we have no reason to

⁸² Remember that being a person means having one's responses imputed on oneself as genuine actions. Accordingly, something as simple as an offer to make a promise is already an implicit way of claiming this status, for such an offer implies a claim to the right to take on commitments.

⁸³ On the other hand, this is also a quite common situation – at least in the context of raising children. Long before they can be held responsible for their responses, in the full sense, children will start to try out actions of “promising” and “keeping promises”, or to implicitly claim other rights that go along with taking on duties. Legal status aside, family, kindergarten and elementary school will (or at least should) provide the sheltered environment where such capacities can be learned, developed and put to test until a child, at about the time it reaches legal age, is concomitantly able to act and refrain from actions, has grasped the concept of responsibility, and is thus able (in principle) to take his or her commitments and duties seriously.

question their personhood. They will be able to keep promises they gave after the traumatic event and, in general, can be held responsible for what they are doing. More challenging is the situation of individuals suffering from anterograde amnesia, who no longer are capable of forming *new* memories. There are tragic cases of this kind where an individual is confined to a time frame of a few minutes. According to what we said above, if someone is lacking the ability to remember his or her deeds, they cannot take any responsibility for what they do. Still, as long as an individual knows what he or she is doing, reflects upon the consequences of his or her deeds for others, and has some understanding of the miserable situation he or she is trapped in, we still have to regard that individual as a person. The last condition (i.e. being able to have an understanding of one's own cognitive limits) is quite important, as persons should be aware of the *extent* to which they can take responsibility for their actions. An amnesiac should know, for instance, what promises he or she is able to keep – and certainly will be very limited in that respect. Anyway, the general requirement of episodic memories is not undermined by these considerations. For one thing, a person who is struck by anterograde amnesia at some point in her life needs to have acquired the concept of responsibility when she was still capable of forming episodic memories.⁸⁴ A child that is losing that capability – say at two years of age – surely could not develop into a person. Furthermore, even though the time frame for keeping new memories may be very limited in anterograde amnesia, it is not nil. If someone indeed *immediately* forgot what he or she was doing right after initiating an action, this would not only prevent them from taking responsibility for their actions, but rather would deprive them totally of their capacity to act.

5.4 Personal Identity and Personality

Since Locke, the question of personal identity has been the question about the criteria of persistence of particular continuants, namely persons. This question is clearly important with respect to morals and jurisdiction. We attribute deeds to persons and so we have to make sure that the person we praise, condemn or convict is the same person that did the deeds we praise, condemn or convict her for. Furthermore, the question is also of special interest because *we* – i.e. the community of persons – are *ourselves* the objects of the investigations. Consequently, issues regarding the criteria of

⁸⁴ Noteworthy enough, in most cases of anterograde amnesia the episodic memories which were acquired *before* the onset of amnesia remain intact. Hence, during the time span that is available to a patient before the next series of short-term memories goes down the drain, he or she can draw on the stock of these unaffected memories for their actions. Needless to say, things look much bleaker for patients suffering from a combination of retrograde and anterograde amnesia.

persistence will be relevant to us so far as we are all interested in our *own* continued survival. (Take for example the question: Is there the possibility of an afterlife?)

Of course, the continuing applicability of the constitutive predicate “person” is a necessary condition for a person’s persistence, but is not at all sufficient to establish that it is the *same* person we’re dealing with at different times. That is why the question of personal identity over time is different from the question of what is a person.

We begin our investigations into the issue of the criteria of personal identity by assessing the most prominent of the so called “physical” persistence-criteria, which were (and still are) favoured by many philosophers. The shortcomings of these accounts will then lead us to what are called “psychological” criteria for personal identity.

5.4.1 Physical Criteria of Personal Identity: Body and Brain

As persons necessarily have to be embodied for them to be identifiable at all (see Section 5.3), one could surmise that they are simply identical with their respective bodies. However, if a person *P* dies, then the constitutive predicate “person” does not apply to its body *B* anymore and hence *P* has ceased to persist. On the other hand, its former body *B* usually still does persist (at least for a while). Therefore persons are *not* identical with their bodies.⁸⁵ However, this observation does not yet rule out bodily continuity as a criterion for *personal identity* over time. Such a criterion could be stated as follows: Person stage *x* at time *t* is a stage of the same person *P* as person stage *y* at time *t*+ Δt if and only if the corresponding body stages *b*(*x*) at time *t* and *b*(*y*) at time *t*+ Δt are stages of the same body *B* according to a suitable criterion of bodily identity (i.e. material continuity). That is, following this proposal, every person persists as the person it is just as long as the predicate “person” applies to a certain re-identifiable body – *its* body. On this account, a person would survive changes of its corresponding body *as long* as these changes do not undermine the criteria for the application of the constitutive predicate “person” to that body. A person could, on the other hand, never survive its body, nor would it be possible for it to persist embodied by another body (“change body”). As a further corollary it would hold: *one* body can only constitute exactly *one* person.⁸⁶

⁸⁵ More or less the same argument also shows that persons are not identical with their *living* bodies, i.e. the (living) *organisms* embodying them (this position is called “animalism”). For if a person *P* has an accident depriving the organism *O* embodying it of the necessary conditions for the application of the constitutive predicate “person”, then *P* has ceased to persist, while *O* may well be still alive.

⁸⁶ Note that this holds to the letter even within the frame of 4-dimensionalism and its assessment of hypothetical cases of fissions or fusions of bodies. While, for example, in a hypothetical fission case two persons could share the same body *stages* before a fission, they would still belong to different *bodies* (which just share stages before the fission point).

This account is attractive to many because of its straightforward simplicity, but the bodily criterion for personal identity still fails.⁸⁷ This is shown by the following thought experiment: Let's assume science had developed to a stage where neurosurgeons would do successful brain transplantations. Let's assume further that, after an accident leaving the brain of a person *P* intact, but damaging the rest of its body *B* beyond any hope of recovery, *P*'s brain is transplanted into a living "donor body" *B'* (say the body of a former person *P'* which suffered brain death). While our strong intuition would be that, under these circumstances, *P* would still persist embodied by *B'*, *P* would have ceased to persist according to the persistence criterion of bodily identity.

Thought experiments like that have led most philosophers who favour a physical criterion for personal identity to drop the criterion of identity of the *body* and adopt a criterion of identity of the *brain* instead: Person stage *x* at time *t* is a stage of the same person *P* as person stage *y* at time *t*+ Δt if and only if the corresponding brain stages *b*(*x*) at time *t* and *b*(*y*) at time *t*+ Δt belong to the same brain *B*.

Unfortunately (perhaps), the criterion of brain identity fails as well. This can be shown by a thought experiment originally devised by Bernard Williams (1973)⁸⁸: Assume we could scan the brain of a person with respect to all the relevant information regarding states and traits of her psyche, like episodic memories, cognitive capabilities and character traits.⁸⁹ Let's further assume that, for purposes of treatment or prevention of certain degenerative processes, we actually could replace (at first only relatively minor) parts of the brain by bio-and/or electronic implants and that we could overcome the problem of preserving the information residing in the parts to be replaced by suitably modifying our implants according to the information derived from a pre-surgery brain-scan.⁹⁰ This would certainly be a most welcome scientific breakthrough. If such interventions involving the replacement of minor parts of the brain proved successful in the sense that all original psychic functions were actually preserved, we would not dream of saying that after the surgery we do not deal

⁸⁷ Psychiatrists may want to raise the objection that the bodily criterion cannot account adequately for the concept of dissociative identity disorder (DID), formerly known as "multiple personality disorder". We can leave this issue undiscussed as the bodily criterion of personal identity also fails for other reasons.

⁸⁸ Actually, we are embellishing the original thought experiment of Williams quite considerably in order to make more obvious and striking its connection and relevance to the topic of new methods of intervening in the psyche.

⁸⁹ It is worth noting that *especially* those who would want to identify the self with the brain should conceive of this as possible at least *in principle* as they would hold that the self is the brain *precisely* because all psychic functions are encoded in the brain.

⁹⁰ This must not be read as meaning that information pertinent to the psyche is stored away in various parts of the brain like dishes in various drawers of a cupboard. Instead, information may "reside in" a part of the brain in the sense that via its mode of connection to other parts the part in question modulates or contributes to the overall information processing in a *specific way*.

with the *same* persons as before the surgery. But if this is so, then there is absolutely no reason for judging differently if such interventions eventually were extended to a point where brain-identity was not preserved anymore.⁹¹ Think, for example, of the successive replacement of considerable parts of the brain by bio- or electronic implants, or by a mix of such implants, while – *ex hypothesi* – preserving all psychic functions. A most radical, empirically improbable, but still conceptually possible case would be the transplantation of a wholly new brain, thoroughly modified to match the results of the brain scan of the original brain. It follows from these considerations that identity of the brain as the criterion for personal identity over time has to be discarded.

5.4.2 Psychological Criteria of Personal Identity I: Memory

The reason why the brain was considered as pertinent to personal identity in the first place is that we (correctly, in all probability) assume that it embodies the psychic functions we are looking for: cognitive abilities and character traits. Accordingly, our investigation so far shows that it is rather these abilities and traits themselves we have to consider, not their contingent physical implementation.

We therefore continue by addressing ourselves to psychological criteria of personal identity. Again, Locke was the first to embrace this approach by suggesting “sameness of consciousness” as the criterion for personal identity over time (see e.g. Locke, 1975:II, 27, §23). Although consciousness encompasses much more, episodic memory is certainly the key to understanding this criterion insofar as consciousness according to Locke involves knowledge of one’s thoughts and actions. “Sameness of consciousness” is constituted by the knowledge of thoughts and actions shared by earlier and later stages of a person via episodic memory. Therefore, even though Locke does not explicitly use the term in presenting his position, he is correctly regarded as trying to give a criterion for personal identity over time in terms of episodic memory. The difficulty is, however, to state the criterion in such a way that certain more or less tricky problems can be avoided. In order to gain insight into the problems to be avoided by an adequate memory-based criterion of personal identity, it is helpful to mention and discuss briefly the main strands of critique with which Locke’s approach was historically confronted.

First, there is the so-called “circularity objection”. It goes back to bishop Joseph Butler and can be stated like this⁹²: A person *P1*’s memory of a past experience cannot be used as a criterion for *P1* being the person *P2* who originally had the experience in question because to establish that *P1* really *remembers* having had the experience, it would already have to be verified independently that *P1*=*P2*. The point is that, according to common usage,

⁹¹ Or at least not according to any non-trivialising criterion for “identity of the brain”.

⁹² In the original writings of Butler, the objection itself is not presented in a very clear way. A contemporary reconstruction can be found in Shoemaker (1970).

“to remember” is a “success-verb”, whose application is correct only if the remembered event really took place. This in turn engenders that, if what is remembered is an *experience*, (i.e. an episode from the first-person perspective), that experience must have been had by *oneself*.

It is quite obvious that the problem encountered here is a mere technicality. To avoid it we need a concept which in all save one aspect behaves like the concept of (episodic) memory: For its correct application it must not presuppose that the episode to be remembered actually happened to the person remembering it. Sidney Shoemaker (1970) coined the term “quasi-memory” for this concept. So, if we could “transplant” episodic memories (or if, in Lamarckian fashion, some episodic memories of parents could be inherited by their offspring), then the persons receiving the memories would quasi-memorise these episodes.⁹³

A second problem with memory-based criteria for personal identity was pointed out by Thomas Reid (1983 [1785]). It draws on a strict reading of Locke’s “sameness of consciousness” criterion, where “sameness” as constituted through knowledge *shared* by earlier and later stages is taken to imply that for various person-stages to belong to the same person the later stages have to faithfully retain the experiential contents of the earlier ones. If succeeding person stages x , y , z are such that x and y share some experiential content E , but z and x do *not* share that content as well, then – according to Locke’s criterion in its strict interpretation – they do not constitute the same consciousness.

With his now famous “paradox of the ‘brave officer’” Reid illustrates that this has quite unwelcome consequences⁹⁴: Imagine a brave officer, who had been flogged for some mischief when he was still a boy, has taken a standard from the enemy during his first campaign as a young officer, and was eventually, at advanced age, promoted to rank of general. Further assume that he still could remember the flogging by the time he took the standard, and that he could remember taking the standard when he was made general, but by then had lost all memory of the flogging. According to Reid, it follows from Locke’s account of personal identity that the boy and the young officer are the same person, as are the young officer and the general, whereas the boy and the general are not – and this would obviously be at odds with the transitivity of the relation of identity.

Reid’s argument has to be disentangled a bit as his presentation is marred by the fact that he confuses continuants with stages. Only because of this the conflict with the transitivity of identity arises. We saw in Section 5.2.4 that different stages cannot be identical anyway, so that questions regarding the identity of continuants encountered at different times have to be put this

⁹³ In fact, if wholesale Lamarckian inheritance of parental episodic memory would be the standard, our understanding of the life of persons would probably be very different (for the result would be a very complex network of psychological fusions and fissions pervading different organisms).

⁹⁴ See Reid (1983):Ch. 6 of “Of Memory”.

way: “Are *O1* at t_1 and *O2* at t_2 both stages of the *same* continuant?” Hence, what Reid *should* have argued is that according to Locke’s theory (in its strict interpretation) the boy and the young officer are both stages of some person *P*, and the young officer and the general are both stages of a person *P'*. But as – again according to Locke’s criterion for sameness of consciousness – *P* and *P'* do not *share* all their stages (the boy-stage belongs to person *P*, but not to *P'*), it follows that $P \neq P'$. And even though this is *not* logically inconsistent, Reid has indeed a point in claiming that this result shows Locke’s criterion of personal identity to be inadequate. For, of course, all three stages – boy, young officer and general – should belong to *one* person.

What this means is that an adequate (quasi-)memory based criterion for personal identity over time has to take into account that persons may *forget*. This is achieved by defining a transitive relation of (*quasi*-)memory continuity.⁹⁵ As this relation is logically complex, the definition has to be broken down into several steps⁹⁶:

(*R**-Def.) Let R^*xy stand in for the following relation: *y* can *quasi-memorialise* at least one episode of experience of *x* from the *x*-perspective (i.e. the first person perspective).

The range of *x* and *y* are person-stages.⁹⁷ As mentioned above, as is the case for memory proper, quasi-memory has to be veridical in the sense that the episodes recalled by person-stage *y* really happened to person-stage *x*, but unlike memory proper, it does not presuppose that *x* and *y* are stages of the same person. It is an indispensable proviso that the episodes are recallable from the first person-perspective of *x*. Else, if I can recall episodes I saw happening to my father, and my father at that time was able to recall episodes he saw happening to *his* father, the intended definition of “memory continuity” building on *R** would yield continuity between the respective stages of my grandfather, my father and me. And as “memory continuity” is supposed to provide a criterion for personal identity, this would eventually lead to the absurd result that there is a person *P* of which these three stages are temporal parts. Note that for the relation *R** to hold between *x* and *y*, *y* does not *actually* have to recall an episode happening to *x*, but just has to be able to do

⁹⁵ Using Locke’s terminology, we could say that “sameness of consciousness” is constituted by *continuity* of knowledge about one’s past experiences rather than by actually *shared* knowledge.

⁹⁶ Usually this is done by employing the apparatus of formal logic to make fully transparent the pertinent structural and inferential properties of the concepts to be defined. However, for the benefit of readers who don’t have a background in formal logic, we just introduce a few symbols standing in for the concepts we are talking about (for the purpose of convenient cross-referencing) and otherwise give the account in plain English.

⁹⁷ As such they have to be embodied. However, having abandoned physical persistence criteria, we do *not* postulate here and in the following that these stages must share the same body or brain.

so.⁹⁸ Note also that, for convenience's sake, we did not attach time-indices to the stages x and y in the above definition. But it is implicit in the definition of R^* that x and y can only stand in the relation R^*xy if it is true for times $t(x)$ and $t(y)$ that $t(x) < t(y)$.⁹⁹ For if $t(x) > t(y)$, then we would not speak of "memory", but rather of something like "precognition". The case $t(x) = t(y)$ is ruled out by the following considerations: Either $x_t = y_t$ or $x_t \neq y_t$. If $x_t \neq y_t$, then these person-stages count as embodiments of different persons anyway (and y 's capacity to "recall" something experienced by x would be considered as some kind of "clairvoyance"). If $x_t = y_t$, then the person-stage in question cannot, strictly speaking, experience something and *at the same time* recall it already.

R^* is irreflexive and asymmetrical.¹⁰⁰ Above all, it is not transitive. We now (recursively) define a transitive relation R using R^* as follows:

⁹⁸ This too must not be read in too strict a sense. For example, while someone is asleep at night he may be considered to be "able" to memorise episodes of the day in the sense that *if* we wake him up and ask him, he will tell us about it. (Generally, this dispositional reading should go for all capacities deemed constitutive for personhood – cf. Quante 2002:20.)

⁹⁹ It is important to note that the times t have to be taken to be (conveniently small) time spans, *not* points in time. The reason for this is that stages have to be spatiotemporal *parts* of continuants. If one successively divides a spatial object into ever smaller parts, the parts will still remain extended. The same has to hold for the division of a continuant into temporal parts. For convenience's sake in what follows we just consider a division of time into "small" stretches of like duration leading to what could be called "standard stages" as smallest units. But of course any temporal part of a person may count as a "stage". This does not at all affect the validity of the arguments in the main text, as long as we disregard stage overlaps or stage inclusions – where the terminology of "later"/"earlier"/"simultaneous" becomes muddled. The ensuing problems could be solved by providing a mereological theory, but more easily they can be avoided by observing the convention that in theoretical contexts the breakdown of continuants into stages should always be carried out such that the resultant stages do not overlap (that is, instead of two overlapping stages we would consider three continuous stages, etc.).

¹⁰⁰ For readers not familiar with the terminology regarding structural properties of relations, we explain at least those types of relations used by us in the text: A two-place relation Rxy is *reflexive* if on the range of objects that can stand in the relation R every object must bear the relation R to itself. For example, every object that at all has a length has the same length as itself. R is *non-reflexive* if there could possibly be objects that do not bear R to itself. E.g. perhaps not everyone loves himself. R is *irreflexive* if no object can stand in the relation R to itself. For example no object can be longer than itself. R is *symmetric* in case that if x stands in the relation R to y , then so will y to x . For example if Fritz is of equal height as Hans, then Hans got to be of equal height as Fritz. R is *non-symmetric* if there can be pairs of objects x, y with Rxy , but not Ryx . For example, Fritz may like Hans, and yet Hans may not like Fritz. R is *asymmetric* in case that if Rxy , then it can never hold the other way round, too. For example, if x is taller than y , then y can't be taller than x . Finally, R is *transitive* just in case that if Rxy and Ryz , then we must also have Rxz . For example if Bill is taller than Jacques, and Jacques is taller than Fritz, then Bill got to be taller than Fritz. R is *non-transitive* if that doesn't have to be the case. For example, if Anna is befriended to Lisa, and Lisa befriended to Wendy, then Anna may or may not be befriended to Wendy. R is *intransitive* if this *can't* be the case. For example, if x is father of y and y is father of z , then x can't be father of z . The way we explained these structural properties, they have to be understood as determined by the *meaning* of the very concepts R in question.

- (R-Def.) Let Rxy stand in for the following relation: The person-stage y is *indirectly linked* via (quasi-)memory to the person-stage x . R is to hold between x and y if and only if there is a finite chain of person stages beginning with x and ending with y where each link is connected to the next through the relation of quasi-memorability R^* .

We can now define the relation of (*quasi-*)memory continuity C :

- (C-Def.) Let Cxy stand in for the following relation: Two person-stages x and y stand in the relation of (*quasi-*)memory continuity. C is to hold between x and y if and only if y is indirectly linked through (quasi-)memory to x (i.e. Rxy), or x is indirectly linked through (quasi-)memory to y (i.e. Ryx), or x and y are the same stage (i.e. $x=y$).

C is a so-called “similarity relation”, characterised by being reflexive and symmetric, yet not transitive. Using the relation C of (*quasi-*)memory continuity, we can now finally proceed to define what it means (within the framework of theories drawing on psychological continuity criteria) for a conglomerate of person-stages to constitute a person:

- (P-Def.) Let Px stand in for the predicate “ x is a *person*”. P is to hold with respect to a conglomerate p of person-stages if and only if
- 1) all person-stages belonging to p as temporal parts are continuous (C) to one another and
 - 2) p is *maximal* in the sense that if some person-stage is continuous to all stages belonging to p as temporal parts, then it is itself a temporal part of p .

From this it follows as a *necessary* criterion of personal identity over time that if two person-stages encountered at different times both belong to some (the same) person, then they must be continuous; and more generally:

- (MemCrit1) *If the members of a set of person-stages M all belong as temporal parts to some (the same) person p , then they must all be continuous to one another.*

Regarding the relation between personal identity and persistence, based on the definition (P-Def.) given above, the question of whether a certain person p persists at a certain time t is rendered trivial: it boils down to the question of whether there is a person stage at t belonging to p .¹⁰¹

¹⁰¹ In line with what we said in Section 2, this definition does not rule out the possibility of “reincarnation”, i.e. that a person persists during a certain time span, then ceases to persist, then persists again later. To exclude this possibility, one would have to incorporate into the definition that for any two stages x and y belonging to p with $t(x) < t(y)$, $t(y)$ is either the direct successor of $t(x)$ (with respect to our division of time into standard durations) or there is a stage z also belonging to p and $t(x) < t(z) < t(y)$.

The memory criterion (MemCrit1) avoids Reid's paradox: the stages of the boy, the officer and the general are not ruled out as belonging to the same person as they indeed are all continuous to one another. Together with certain assumptions¹⁰², (P-Def.) also yields directly a *sufficient* criterion for personal identity over time in the sense of:

(MemCrit2) *If the members of a set of person-stages M are all continuous to one another, then there is a person p which they all belong to as temporal parts.*

Note that a special case of this is that every person stage belongs to at least one person.

So much about theory. Let's now see whether memory continuity is really a necessary criterion for personal identity. There are two classic thought experiments – one by Locke and one by Leibniz – which make it extremely plausible that this is indeed the case.

Locke's thought experiment was originally designed to refute the so called "simple view" of personal identity, which says that personal identity is constituted through sameness of soul, where "soul" is understood to be a simple, indivisible and immaterial substance.¹⁰³ Apart from that, the thought experiment makes a good case for memory-continuity as a necessary criterion for personal identity. Locke (1975:II, 27, §14) invites us to assume that some present-day person is the reincarnation of Nestor at the siege of Troy in the sense that the present-day person shares with Nestor the same "immaterial spirit" (thinking substance), but has no consciousness whatsoever of any of

¹⁰² The first assumption we need is trivial, namely that for all sets of person-stages *M* the corresponding conglomerate *p* of person-stages exists. This can be secured by the simple decision to regard every possible collection of stages as a "conglomerate". The second assumption is more demanding as it says that all sets of stages where all members are continuous to one another are *finite*. This amounts to saying that no one persists forever or has persisted ever since. Both cases cannot (or should not) be ruled out on conceptual grounds alone. However, the second case at least can be excluded for obvious empirical reasons, and there is actually no need to worry about the first case, considering that – since we do not know the future – the constitution of continuants is *de facto* *always* only done "up until now". While the definition (P-Def.) provides a test regarding the question of whether a given conglomerate *p* satisfies the definiens, it does not give any hint concerning the question of how to construe such a conglomerate. This can be achieved the following way: Start with one stage *a*. Then pick another stage *b* that is continuous to *a*. Then pick a third stage *c* continuous to both *a* and *b*, and so on, until there is no stage left that is continuous to all stages already picked (the second assumption secures that this result can always be achieved in a finite number of steps). What we get is a conglomerate *p* satisfying definition (P-Def.). Fortunately, this way to construe persons from the set of stages is not something we do (or should do) in real life. We only use it as an easy way to see that (MemCrit2) *indeed* follows from (P-Def.).

¹⁰³ Obviously, this is the orthodox view within Christian theology. It is still very influential today and, in recent times, was stoutly defended by philosopher Richard Swinburne (see e.g. Shoemaker and Swinburne 1984).

Nestor's actions. Would we deem the present-day person the *same* person as Nestor? No more, Locke says, than if some (or even all) particles of matter that once were part of Nestor accidentally were now part of the present-day person: Sameness of *immaterial* substance without "sameness of consciousness" will no more make for the same person than sameness of *material* substance without sameness of consciousness would. And indeed, if mere sameness of material substance will not suffice to constitute personal identity, then it is not intelligible why resorting to immaterial substances should make any difference. What counts in personal identity is *never* sameness of substance, be it material or immaterial, but "sameness of consciousness" (which we have to conceive as "memory continuity").¹⁰⁴

Leibniz's thought experiment – which he devises in "Discourse de Métaphysique", Section xxxiv – is similar to Locke's, but probably even more convincing as it is placed in the context of a *choice* everyone can imagine to be presented with. Suppose someone could instantly become King of China (which Leibniz obviously assumed to be a most pleasant thing), but on the condition that he would forget everything he had ever been and experienced: as if he had been born again. Leibniz then considers the question whether it would be rational for a person presented with this option to take it – and unhesitatingly answers in the negative. For this would, he argues, be indistinguishable from that person being annihilated and a King of China being created at the same instant in her place. That is, taking the option would amount to committing suicide – not in the sense of ending one's life as a human being, but as the person one had been.¹⁰⁵

So far as the assessments of Locke and Leibniz are in line with the intuitions of most people, their thought experiments do show that the criterion of (episodic) memory continuity states a necessary condition for the attribution of personal identity. Still, the criterion also allows for a good deal of "robustness" of personal identity with respect to memory loss: The case of Alzheimer-patients provides evidence that in the course of a degenerative

¹⁰⁴ If this is so, it can easily be seen that sameness of immaterial substance cannot even lay claim to being a necessary condition for personal identity. For why shouldn't immaterial substances successively pass through a person during the time of her persistence just as material particles pass through an organism during its life? In both cases, as long as the right "structure" is retained, everything should be just fine. Hence, as the assumption of immaterial substances obviously does not help at all in accounting for personal identity, we encourage the reader to regard this fact as a sufficient refutation of the "simple view". Even though it may be consistent with all empirical data, the assumption of there being immaterial substances should be dropped in accordance with "Occam's razor": "Entia non sunt multiplicanda praeter necessitatem."

¹⁰⁵ Like Locke, Leibniz also adopts the classical position that "person" is a moral category marking individuals who can be held responsible for their actions. In this vein he says (in the same section of the "Discours de Métaphysique") that "it is the memory or knowledge of this 'I' which renders it open to punishment and reward".

process persons can persist in spite of *considerable* loss of episodic memory. In this, it is an important stabilising factor that their character-traits remain suitably stable for their social contacts to “recognise” and anticipate their “typical” actions and behaviour. However, this “robustness” of personal identity does not extend as far as to cover also the final stages of the disease where episodic memory is eventually too attenuated and disintegrated to still provide for suitable interaction – think of the stages when the patients don’t recognise their relatives and friends anymore.

5.4.3 Psychological Criteria of Personal Identity II: Personality/Character

Now the crucial question is: Is (episodic) memory continuity not only necessary, but also sufficient for personal identity over time?¹⁰⁶ Instances of radical personality change, as most strikingly exemplified by the fictional character(s) “Dr. Jekyll and Mr. Hyde”, render this doubtful: In Stevenson’s novel, episodic memory seems to remain unaffected through the back and forth transformations between Jekyll and Hyde. After all, Dr. Jekyll at least initially takes delight in living out his “dark side”, which could hardly be explained had he forgotten everything he did as Mr. Hyde. Still, it seems natural enough to suggest that we are faced in the story with altogether different *persons* rather than with one person repeatedly undergoing *personality changes*. Admittedly, it is by no means crystal clear how this case and related – less dramatic, but non-fictional – cases are to be judged.¹⁰⁷ But, tellingly, our intuitions are not as unambiguous as in the thought-experiments destined to explore whether memory continuity is *necessary* for personal identity. We do not say: “Well, *of course* they must be one and the same person, for they stand in the relation of memory continuity!”

¹⁰⁶ In some way it is *certainly* not a sufficient criterion because it is stated much too liberally. The antecedent of MemCrit2 could already be fulfilled if there is just one single isolated quasi-memory connecting me to someone else. However, this is not a decisive argument against memory continuity *as such*. It merely shows that, to be sufficient, the relation *R* would still have to be strengthened. Doing this by adding constraints on the *amount* of quasi-memories or their causal *origin* (as promoted by some philosophers) is not very promising. Rather, it is their *content* that needs to be further specified. Instead of being isolated entities, memories usually are connected to other memories, so that each of them has a distinctive “place” within the network of one’s memories as a whole. If the quasi-memories of a person-stage *y* connecting it with the experiences of a person-stage *x* do not “fit in” with the whole of *y*’s network of memories, they may not offer sufficient evidence for establishing that there is a person the stages *x* and *y* both belong to. However, we will not attempt to specify the criterion of content-connectedness of memories any further here. The reason is twofold: firstly, we will see that a criterion of memory-continuity cannot be sufficient for personal identity anyway, and secondly, the sort of connectedness we just referred to is much better explained within the “narrative account” of personal identity, which we will expound later.

¹⁰⁷ As a convenient litmus test for our intuitions we may ask ourselves whether we would tend to hold Dr. Jekyll responsible for Mr. Hyde’s deeds. If they are the same person, we certainly should do so.

So, when it comes to the question what is *sufficient* for personal identity, then, in addition to memory-continuity, personality-traits seem to be relevant, too. As additional evidence for this observation one may cite the common way of talking about the “identity” of persons, in which the term is more or less synonymous with, or at least closely related to what is called “personality”.¹⁰⁸ This would be not at all mere coincidence if (aspects of) personality should turn out to contribute in a pertinent way to questions regarding the persistence of persons. However, in order to see how exactly we may refer to it in framing criteria for personal identity, we need to analyse the notion of personality in somewhat more detail.

What ever personality is, it surely is not to be understood as a simple property. Therefore, we first try to analyse the complex bundle of interacting traits (p-traits, for short) personality seems to consist of. To get a grip on the question as to what p-traits are and what distinguishes them from other traits persons may have, we set out by defining the comparatively unsophisticated concept of “habit”: Habits are dispositions to behave or act in a certain way in certain types of situations. Dispositions themselves may be established and disestablished again, and there is no constraint on how long a disposition can or must hold to count as one. But habits not only are “relatively stable” – that is the underlying associations between types of situation and types of action or behaviour have to hold over a suitable length of time –, but they also have to manifest themselves suitably often.¹⁰⁹ Habits are not themselves p-traits, however: e.g. the habit to brush one’s teeth, drink coffee in the morning or raise one’s hat to greet people one encounters would not count as p-traits (or character-traits). A single p-trait is not just a stable disposition to act/behave, but rather comprises a wide range of dispositions of a certain more general kind. Instead of trying to define this more precisely, we settle for giving examples: For instance, the p-trait of “friendliness” will comprise habits like greeting people, giving polite answers to questions, readily offering help when needed etc. Or the p-trait of “being conservative” will comprise various dispositions favouring established or well-tried situations, procedures and institutions over new alternatives. It is far from easy to categorise such traits. According to the well-

¹⁰⁸ We already alluded to this usage that is sometimes picked up by psychology too when discussing the case of McMurphy in the introductory section where we found that, in this way of talking about “identity”, more emphasis is given to those personality traits which make the person in question differ from other persons, i.e. those which – in their combination – make him “unique”.

¹⁰⁹ What length of time and what frequency of occurrence are to be regarded “suitable” depends on the context at hand. A mayfly during its short life might be said to have had the habit to fly in the direction of light-sources. On the other hand, I probably can’t be said to have had the “habit” of coming home late last week, or throughout my life to have had the habit of visiting my parents on turns of centuries. It seems to be futile here to try being much more precise than that.

known classification¹¹⁰ by Joy Paul Guilford, we would have the following categories: temper, attitude, interests, needs, abilities (talents), morphological traits and physiological traits. Out of these we would dismiss morphological traits (e.g. skin-colour) and physiological traits (e.g. being prone to sweat) right away as it seems not to be adequate to regard them as p-traits in a strict sense (even though they may under certain circumstances very well causally condition, interact with or “favour” certain p-traits in the more strict sense).

Temper is the oldest known category of p-traits. Temper comprises long-term emotional dispositions.¹¹¹ Hippocrates – the “father of medicine” – sorted these into typical bundles, constituting the sanguineous, choleric, phlegmatic and melancholic personality types. Based on similar considerations by Wilhelm Wundt (1903), Hans J. Eysenck (1965) postulated two dimensions – introverted/extroverted and stable/unstable – in which all traits belonging to temper are supposed to be locatable.¹¹² If the dimensions are graphically represented as orthogonal lines defining a plane, the four sections comprise temper-traits belonging to the classical four tempers distinguished by Hippocrates: introverted-stable=phlegmatic, introverted-unstable=melancholic, extroverted-stable=sanguineous, extroverted-unstable=choleric. Accordingly, “depressed” (German: “bedrückt”) – the typical attribute of the melancholic type – is about equally accounted for by “introversion” and “instability”. On the other hand, “anxious” and “unsociable” also belong to the melancholic personality, but “anxious” loads higher on instability than on introversion, while for “unsociable” it’s the other way around.

¹¹⁰ There are other classifications, of course. We especially mention Guilford’s because it is not only one of the most prominent classifications, and probably the most exhaustive, it is also the easiest to start from, with the goal of conceptual clarification in mind.

¹¹¹ One may be in an aggressive *mood* without having an aggressive *temper*. One could say that the temper of an individual is characterised by those moods – relatively short term dispositions to feel (and thus act) in a certain way – which are “chronic” in that individual, that is, occur so often and/or last so long as to pervade the life of the individual, shaping the whole of its behaviour and agency.

¹¹² For the reader to grasp Eysenck’s model, we have to explain briefly the basic features of so called “factor analysis”. Factor analysis is a mathematical method (developed by the psychologist Charles Spearman at the beginning of the twentieth century) for grouping variables according to their mutual correlations. But instead of just putting the variables into *categories*, new variables – “factors” – are abstracted (“extracted”) from the correlation-data, and then used to describe the original variables and their relations: For every variable and factor, there is the measure of how much the variable is “loaded” with the factor. The higher the loading, the higher the contribution of the factor to the value of the variable. The goals are to find factors which are “orthogonal”, i.e. uncorrelated to one another, to have as few factors as possible that still allow for a most exhaustive description of the relations among the original variables, and – last but not least – to keep the factors easily interpretable. Eysenck’s dimensions “introverted/extroverted” and “stable/unstable” are just such orthogonal factors.

By his category, “needs”, Guilford does not address “universal needs” like the need of food or fluid intake for survival, as such needs do not differ in any interesting way between persons.¹¹³ Rather, the traits he has in mind would belong to a broader category comprising *basic long-term motifs* (interests, goals, values) which persons may or may not have, or may have in varying degrees. Guilford’s “interests” category (to be understood in the common way of “taking an interest in specific activities” – such as “thinking”, “chatting”, “doing needlework”, “climbing mountains”, etc.) would also fall under that broad category as long as they are long-term and basic in the sense that they are executed for their own sake (not just as means to *other* goals).

“Attitudes” towards all kinds of issues (things, facts, proposals, norms etc.), while quite correctly being associated with personality, at first may not seem to constitute an independent group insofar as they are analytically bound to motifs via mediation of one’s (long-term) beliefs: If my goals include getting good education for my offspring, and my beliefs include that college *C* is sub-standard, then – all other things being equal¹¹⁴ – my attitude regarding the idea to have my daughter attend college *C* is already fixed. However, beliefs are *themselves* (propositional) attitudes. And while certain beliefs may be “favoured” by one’s motifs, and certain motifs by one’s beliefs, no group can be reduced to the other. Hence, *basic* (=non-derivative) *long-term propositional attitudes* indeed form a category of their own that in conjunction with the category of motifs determines a person’s attitudes in the wider sense.

Finally, what about *talents*? A talent for *X* is an innate¹¹⁵ and, compared to the average, increased capacity¹¹⁶ to learn whatever is required to successfully perform the kinds of actions related to *X*. Are talents p-traits? Or do they just interact with the personality?¹¹⁷ Considering the common use of “personal-

¹¹³ When inquiring about someone’s *personality*, we would find it rather strange to be informed that the person in question has got to eat, drink and sleep on a regular basis.

¹¹⁴ Meaning that in the complex net of purposes and beliefs no other purpose-belief connections serve to outweigh or override this connection.

¹¹⁵ Although one can indeed acquire talents – by maturing –, and though one may be fostered in this process of maturing, it would be wrong to claim that talents themselves can be “learned”. By which actions could someone learn not merely how to play the violin well, but rather the *talent*, i.e. the increased capacity to learn to play the violin well?

¹¹⁶ “Increased capacity” does first and foremost mean that the achievable *level of success* is higher than in non-talented competitors. Though this does most often include the ability to achieve levels *faster*, this ability is usually not to be identified with the talent itself. To be sure, someone *may* be said to have the “talent” to get into the very basics of certain things extremely fast (and then be forever stuck there). In that case, however, the description of the respective talent would have to mention that, explicitly.

¹¹⁷ There are certain usages of the word “character” by which talents would not be regarded as part of the character, namely when “character” is restricted to such traits – like “reliability” or “greed” – that are subject to moral judgements. This usage comes to the fore in phrases like “she has character”, “he’s got a bad charac-

ity” it seems that talents do not just interact with, but rather are to be considered an integral part of the personality of a person. For example, could one adequately describe the personality of Einstein without mentioning his knack for physics? But then again, what about latent and unused talents? Could it be that only fully developed talents form a part of one’s personality? We think this should be answered to the negative, at least if one does not want to exclude a person’s *potentials* from the p-traits of that person. A person may never have gone into politics, but he or she nevertheless may have had all the prerequisites of becoming a great social reformer (and if this does not count as information about personality, then what does?).¹¹⁸

In a nutshell, our analysis of personality in terms of the character-traits it seems to consist of yields the result that personality is the system of interacting *temper*-traits, (long-term) *motifs*, (long-term) *propositional attitudes* and *talents*. Regarding the quest for a not only necessary, but also sufficient condition for personal identity over time, one could now try to augment the requirement of memory continuity by postulating that, additionally, a certain yet to be specified relation pertaining to personality (character) has to hold between two person stages for them to belong to one (the same) person.

It is quite easy to see that a mere relation of “character continuity” (defined in formal analogy to memory continuity¹¹⁹) will not do the job: Just compare the logically possible case where – for whatever reason – a character-change occurs continuously, but radically, in a matter of days (or maybe hours), with the case that – somewhere within the same time period – an equivalent result is brought about “abruptly” (non-continuously). Whatever we would decide regarding the question of personal identity in such cases,

ter” etc. But if we consider that traits of temper *definitely* belong to character, this already goes to show that a purely moral usage of “character” is too narrow, as certainly many temper-traits are neutral in that regard – e.g. to be anxious is *per se* neither a virtue nor a vice.

¹¹⁸ However, it is worth noting that from the inclusion of talents into the range of p-traits it does not follow that the concrete abilities (psychic or physical) one acquires in using a talent are themselves p-traits. This would indeed stretch the use of “personality” or “character” beyond what is adequate. A person may have the talent to do maths, but only that talent, not the particular ability to solve quadratic equations is part of her personality.

¹¹⁹ One would start out with a basic relation of “(sufficient) character overlap”, where two stages have (a certain minimum amount of) p-traits in common. One then would go on defining that two person stages stand in the relation of “character continuity” if they are indirectly linked via a chain of stages where each two succeeding stages stand in the relation of “(sufficient) character overlap”. Eventually, one would arrive at a definition stating that a conglomerate of person-stages constitutes a person if and only if the stages all stand in the relations of (quasi-)memory continuity *and* character continuity to one another, and the conglomerate is maximal. According to such a definition, each relation would provide a necessary criterion for personal identity, while together they would (purport to) yield a sufficient criterion.

we would most probably not treat the two cases *differently*; and even *if* we did so, then definitely not *because* the change was “continuous” in the first case and “non-continuous” in the second.

A lesson which, one may think, has to be learned from this is that a character-based criterion has to draw on a demand for a certain “stability” of a person’s character over time rather than on continuity. On the other hand, it is clear from the outset that at least the most straightforward form of such a demand – namely that a person’s character should *not at all* change – would be blatantly false (pace Schopenhauer). Hence, even while stability of character and (quasi-)memory continuity combined would maybe *suffice* for personal identity over time, this is achieved only on pain of that combined condition not being *necessary* anymore because one ingredient (character-stability) is not necessary for a person’s persistence. Nothing is more obvious than the observation that not every change in personality already amounts to identity change.¹²⁰ For good or bad, the personality of persons is bound to develop throughout their lives. Hence, an adequate criterion for personal identity over time must leave room for personality changes. More than that: it is also futile to single out any specific *subset* of p-traits from the categories *temper*, *motifs*, *attitudes* and *talents* which is such that at least *its* members have to remain stable for a person to “survive” as the person she is. For there is no such set: All else being equal, any given p-trait may at some point during a person’s life be subject to change without the person ceasing to persist.¹²¹

So what are we to do with this? As a last try one could seek to reconcile the conflicting demands for continuity and stability by proposing a criterion of

¹²⁰ While this statement hardly seems to require justification, let us give one reason (just in case): If a person’s character could – for conceptual reasons – never change, then any observed change in personality would imply that a person (the one who used to be there prior to change) has actually ceased to persist, giving way to another, the “post-change” person (both being successively embodied by the same organism). This in turn would imply that one could, for example, not be convicted for a crime after a personality change, for the crime would have been committed by someone else entirely (namely the pre-change person) – which of course is absurd.

¹²¹ What if one suggested that, while every single p-trait may indeed change, the character “as a whole” must not change beyond a certain degree (say 50% of the p-traits) for still being able to talk about “the same” person? This is, of course, totally out of the question as it is open to “Reid”-type objections: Imagine our brave officer changing about 25% of his character-traits from boyhood to young adulthood, and then again such that in the end somewhat more than 50% of the p-traits have changed along the way from boyhood to old age. Again we would have the undesirable case that boy and young officer constitute one person, and young officer and general *another* – with the young officer belonging to *both*. But this is not where it ends: Definitions in this vein could split up one person into a whole bunch of “overlapping” persons, depending on *where* one cuts out phases just satisfying the criterion (looking continuously back in time from the old age undoes certain changes, which in turn gives room for taking in more of the boyhood personality – and *vice versa*).

character continuity with “relatively slow” overall character change – calling this “relative character stability”. – But this would be intolerably vague, if not really desperate. How leisurely paced would we like character change to proceed so as not to endanger personal identity? We confess that we cannot see any non-arbitrary answer to this. Even *if* we would resort to just setting certain constraints by decision (of the sort that there must not be more than n p-trait changes for two person stages that are less than a certain stretch of time Δt apart), this would not help at all. The reason is that every such notion of “relative character stability” would still have a continuity relation built into it. With reference to that “ingredient”, it can always be challenged by imagining some Methuselah who fulfils the relevant criteria for being a person all through his age-long life, but – while all stages satisfy the combined criteria of (quasi-)memory continuity *and* relative character stability (whichever way the latter may be spelled out) – the late stages bear *no* direct relation, neither memory- nor character-wise, to the early stages. Wouldn’t we be inclined to say that in such a case the early stages and the late stages constitute different persons? In fact we do not have to decide this question to see the inadequacy of “relative character stability”. It suffices to note that the Methuselah case is setting the stage for the very same objection we’ve already discussed above with respect to the relation of “mere” character continuity: If we compare two “Methuselahs”, where in both cases the late stages are radically different from the early ones, we would not dream of treating them differently with respect to questions pertaining to personal identity (like whether we could hold the old men responsible for crimes they committed during their youth) just because we were told that in one case the intermediary stages all satisfy the requirements of “relative character stability”, while in the other case there had occurred a more “rapid” character-change somewhere along the way.

The message to be taken from this discussion is that psychological continuity relations – whether drawing on memory, p-traits or a combination of both – will *never* provide sufficient criteria for personal identity, but at best necessary ones (like quasimemory continuity).¹²² On the other hand, any strengthening of (quasi-)memory continuity by way of combining it with some strict psychological equivalence relation¹²³ (like “sameness of charac-

¹²² Indeed, in its original context, the “Methuselah” example (as discussed by Lewis 1983: “Survival and Identity”, IV) does not deal with character, and rather should be viewed as arguing for the inadequacy of any mere continuity relation for providing a sufficient criterion of personal identity. We therefore could have used the “Methuselah” example already above to support the argument that (quasi-)memory continuity is by itself not a sufficient criterion for personal identity. The reason that we didn’t do so, but rather drew attention to the possible overriding force of radical character change, is that the full strength of the “Methuselah” example is only exploited if it is used to show that even (quasi-)memory continuity and character continuity combined do not yield a sufficient criterion.

¹²³ Equivalence relations are reflexive, symmetric and transitive.

ter”) will *at best* provide a sufficient criterion, however on pain of not being a necessary one anymore.

5.4.4 A Narrative Approach to Personality

Alas, it is time to admit defeat on this particular battlefield and look out for a whole new tack on the issue. What we tried to do until now was to specify necessary and/or sufficient conditions that persons need to fulfil in order to remain who they are over a course of time. The failure of all these attempts could be taken to indicate that, as a matter of principle, this type of criteriological approach will not yield a satisfying account of personal identity. This is because its feasibility critically hinges on the assumption that problems of personal persistence may be settled by referring to changes in psychic traits that can be qualified without further reference to the specific *situation* in which these changes occur. In this way, we feel entitled to say, for example, that *whenever* a being suffers a total loss of its language abilities it cannot be considered a person afterwards. However, there is a conspicuous difference between situations in which a being’s person-status is in question and situations in which we come to decide cases of identity change. While in the former situations the affected being will hardly “raise its voice” on its own behalf¹²⁴, in the latter cases we are considering persons that have a stake in their being recognised as still or no longer the same and as such will themselves *participate* in the process of deciding about the pertinence of the kind of change they underwent. Maybe it is due to its participatory nature that the social process that mediates our mutual recognition as one-and-the-same in different situations cannot be adequately analysed in terms of context-invariant criteria. To further investigate the plausibility of this reasoning, our next step will be to develop a theoretical account of the process of challenging and defending claims for or against personal persistence that promises to do better justice to its inherent context-sensitivity.

In order to attain a reasonable distinction between gradual personality changes, on the one hand, and outright changes of personal identity, on the other, we first need to establish a more adequate understanding of personality. Up to now we treated personality more or less as a mere aggregate of p-traits. This account would be adequate with regard to the character of animals or pre-personal toddlers. In the previous section we stated that many character traits can be ascribed to non-personal beings, provided they display suitably complex behaviour with stable long-term dispositions. Their *character* is nothing over and above the sum total of these ascriptions of character traits. Still, the difference with regard to *personality* is not just that next to these character traits persons also possess some further (or, for that

¹²⁴ If it does, this will of course suffice to falsify the hypothesis that it is not a person we are dealing with. Please also keep in mind what we said in the previous section about personhood, namely that having “language abilities” does not presuppose the ability to *speak* (i.e. using the larynx).

matter, even *many*) p-traits that cannot be ascribed to beings that aren't persons because of them lacking certain abilities. In this vein it would be right to say that a person, but not an animal can be virtuous or courteous as these traits require a conceptual grasp of certain values or norms that is beyond reach for any known animal mind. However, compared to character as it may be attributed to beings which are not persons, personality is not merely a construct with some extra dimensions. Rather, there is a qualitative difference due to the fact that persons are self-referential beings, who not only *have* character traits, but, moreover, hold a set of beliefs *about* these traits as well.¹²⁵

What persons believe about their own p-traits is part of a more complex system of self-related beliefs (and, as we will see, motifs) that constitutes their *self-concept*. When, in the third section, we found that persons as moral agents need to have a concept of *themselves* being members of a community of "others like them", we brought to bear the concept of a self in a most general form. Actually, all individual persons at any given time of their persistence entertain a much more colourful self-concept, which they may – under inviting circumstances – disclose by numerous statements involving the personal pronoun "I". However, by no means all self-related statements constitute part and parcel of a person's self-concept. For instance, I do not reveal anything about how I understand myself by stating "I had a bad night's sleep." Just like personality, the self-concept does not represent how a person is at the present moment or how she thinks about herself today, but rather it offers a more general or long-term account. Tentatively, we can single out the relevant statements by postulating that they may be uttered in response to the question "Who are you?" Of course, when confronted with this somewhat dubious question, we usually make do with saying our name. But suppose our opponent is not satisfied with our reply, and provided we take a sincere interest in presenting a comprehensible picture of who we are, then we may start telling a more detailed story including distinguishing features of various kinds: As said before, we may express opinions on our p-traits. Given the way we introduced these traits, this includes statements about our values, preferences, and what we appreciate in life ("I hate injustice", "I am a soccer maniac" etc.). However, a person's self-concept is not exhausted by beliefs on his or her p-traits. For instance, while we excluded morphological and physiological traits in our review of Guilford's account of personality, *beliefs* about traits of these kinds may very well form part of personal self-concepts ("I am a particularly handsome and good looking guy", "The root cause of all my problems with others is that I am not attractive", etc.). Even though we took pains to conceptually discern persons from their bodies, it cannot be neglected that in talking about themselves persons frequently refer

¹²⁵ This is (or anyway should be taken as) already implicit in that part of Locke's definition of "person" according to which a person can *consider itself as itself*. If a person can do this, she will necessarily have what we will call a "self-concept" and expound in the paragraphs to follow.

to their bodies. Furthermore, in the stories they tell about who they are, persons like to include pieces of information that relate to their social position by specifying e.g. their occupation or certain of their achievements (“I am a member of parliament”, “I am a father of six children” etc.). Finally and most peculiarly, who a person is in part determined by who he or she *wants* to be.¹²⁶ That is, a person’s self-concept includes her projects of “self-creation”.¹²⁷ In summary, we don’t want to be very restrictive regarding the question as to what kinds of self-related beliefs persons may incorporate in their self-concepts. Everything persons consider to be “typical” or significant for their individual “way of being” can be a proper part of their self-concept.

Statements by which persons seem to express their self-concepts can be untrue in two different ways: Firstly, as a matter of fact, persons at times deliberately deceive others about their personality. For various reasons (too obvious to mention), people sometimes want others to believe that they have certain p-traits which they actually lack. Usually, when persons deliberately give false information about their personality, they will (at those occasions) also try to act accordingly, so that those they are trying to cheat will more readily believe that they really are the kind of person they claim to be. Of course, it would be wrong to say that persons express their self-concept by statements that are uttered to deceive others. A person’s self-concept is constituted only by those beliefs that he or she honestly considers to be true. By deceitful self-related statements and by those actions that should render them credible persons actively create *images* of themselves.¹²⁸ It seems quite safe to claim that, next to their self-concept, most persons entertain one or more images fitting different purposes and contexts. Now, to unmask images, it is not enough to find out that the information persons convey about their personality is not correct – after all they may honestly believe

¹²⁶ This is what in psychology is called the “ideal self-(image)”. My ideal self does not consist in how I see myself, but rather how I *want* to be or *become*. It could be said that one’s ideal self is the *normative dimension* of one’s self-concept.

¹²⁷ This particular aspect of personhood has been analysed with great care by Jonathan Glover (1988).

¹²⁸ Since the dichotomy of self-concept versus image is of paramount importance for psychotherapy, numerous terms in different languages (that frequently are not readily translatable) have been introduced to address it. In English, the term “persona(e)” is sometimes used to refer to the social role(s) a person may play. However, we decided for the more neutral term “image” in order to avoid the baggage “persona” carries due to its origin in archetypal psychology. To avoid possible misreading of our discussion about “images” in the main text above, it is in order, however, to point out some more things pertaining to ideal self-images: Just as the ideal self-image does not describe how I actually see myself, it also must not be confused with how I want others to see me – in those mere “images” I may create of myself. Persons also try to act according to their ideal self-image, just like in the case of a mere “image”. However, in doing this, they do normally *not* want to deceive others about their p-traits, but just to attain these traits as habits. In fact, people are often perfectly *honest* about these things (“I always try to be friendly, but once I’m under stress, I tend to lose my temper”).

that they really are as they are claiming to be. So how can one know what is genuine self-concept and what is image in these cases? Well, as the saying goes, you can fool some people all the time, and you can fool all people some of the time, but you can't fool all people all the time. In contrast to acting in accord with one's p-traits, pretending needs some amount of effort. It is not quite that easy to live up to other's various expectations by compromising one's genuine self-concept without anyone ever noticing. If a person tries to convey the information that he or she has a particular p-trait in a certain type of situation and to certain other people only, such that at other occasions she does not care to act as if she herself believed she got the trait in question – then that is the sure sign of an image.¹²⁹

The second way in which self-related statements can go wrong is that persons may simply err with regard to who they are. Someone may think he is quite self-composed, yet, his short-tempered way of reacting may reveal his choleric temper. Numerous psychological studies have shown that we are by no means completely transparent to ourselves. Much to the contrary, we can be mistaken with respect to almost each and every aspect of ourselves. So, even though only honest convictions should be considered part of a person's self-concept, it can (and usually will) contain false beliefs as well as true ones.

Self-related beliefs are false if they do not appropriately represent the way the person is who holds them. In other words: false components of a person's self-concept do not match with her *personality*. Whenever someone offers an account of another person's personality, they basically try to answer the same question we found to be constitutive for personal self-concepts, namely *who that particular person is* compared to other persons. Self-concepts represent subjective points of view on personalities. Their subjectivity, however, lies not in their being independent from other persons' points of view on the same object. Much to the contrary, the way persons come to think about their personality is deeply influenced by what others think about who they are.¹³⁰ Rather, a self-concept is subjective in the sense that it comprises *the person's* thinking about what marks him or her off from other persons, no matter whether it truly represents the person's personality or not. Consider, for instance, an anorexic young woman who is utterly convinced to be too fat even though she is in fact close to starving. It is a genuine part of her self-concept that she considers herself to be too fat. However, in giving a faithful account of her personality one cannot refer to her being (too) fat, for the simple reason that she is not. All we can say with respect to an anorexic woman's personality is that it is a salient feature of her that she holds

¹²⁹ Of course, sometimes, especially when confronted with inconsistencies between self-related statements and their behaviour, persons may even openly admit that their words or deeds are founded in a consciously designed image.

¹³⁰ Actually, an important way of learning words for p-traits is to understand how others (most prominently parents) respond to one's own behaviour by the ascription of p-traits (e.g. "You are a naughty boy"). In this way, others already exert an influence on a person's self-concept before she acquires the means to make up her own mind about who she is.

wrong beliefs with regard to her outward appearance, or that she has got a disturbed perception of her own body.

Another way to misunderstand the notion of subjectivity in relation to personal self-concepts would be to think that nobody but each individual person is in a position to know what makes up the contents of her self-concept. Most intuitions indicating a particular *epistemic authority* of persons regarding their self-concepts are based on the conviction that it is a *private affair* how a person thinks about herself¹³¹. In order to see that this is not quite right, just imagine a human resources manager who publicly purports to have a decent egalitarian attitude with respect to racial differences. Suppose he is so cautious not to reveal his prejudice against black people that he will never communicate it to anyone. However, in the privacy of his office he visits racist websites and enjoys laughing at jokes mocking black people. Now *if* this behaviour got revealed beyond doubt (for instance by a hidden camera in his office), this would be sufficient evidence to conclude that having racial prejudices is not only one of his p-traits, but indeed also a part of his own self-concept. This diagnosis could even prove robust against any later efforts of his to hold up his egalitarian image by trying to “explain away” the evidence in question once he got confronted with it (“I was just trying to get information in order to better be able to fight racism” etc.). What the example shows is that there can be good reason to attribute beliefs to a person’s self-concept even if that person consistently denies holding them. Of course, if the example is found to be convincing, then this is because a person simply cannot *avoid* being/becoming aware of his holding of racial prejudices if he really spends a good deal of time in active search for racist jokes which he then finds himself smirking, giggling or outright laughing about. To be sure, it has been established in a number of studies, employing *inter alia* “Implicit Association Tests”, that white personnel managers may be unconsciously biased against black applicants.¹³² However, unconscious bias needs to be distinguished from covert (but nevertheless conscious) prejudice, and only the *latter* can be part of personal self-concepts the way we want to conceive of them here.¹³³ Of course, aspects of the self-

¹³¹ Private in the sense that, as a matter of principle, only I myself have access to and, hence, can really *know* about it – while others always have to rely on my information.

¹³² The classical experiment has been conducted by Word et al. (1974). For a host of further references see the “Project Implicit” website at <http://projectimplicit.net>, accessed on December 8th, 2006.

¹³³ Which is: as the actual “self-understandings” of persons, as the ways persons actually do “consider themselves as themselves”, and *not* as some part of their psyche that – however important – is unavailable or intransparent to them. What e.g. a psychoanalyst does is (in a first step) to help people in *becoming aware* of repressed beliefs, goals, conflicts etc. and thus to help them in descriptively *adjusting* (not discovering!) their self-concept. Only on the basis of such “insights” on the part of the clients, further progress can be made towards resolving psychic conflicts and any symptoms possibly connected to them. This will most often involve a change of the ideal self-image of the client, too.

concept a person entertains may very well remain private in the harmless sense that they will *de facto* never be known by anyone else (maybe the covert racism of the manager really goes unnoticed). What doesn't make sense, though, are privacy assumptions along the line that – “as a matter of principle” – no one but the person in question can *ever* know about aspects of her self-concept.

What we just dealt with (and refuted) was the claim that a self-concept is a private entity, one which only the person who actually has it has access to, and that therefore persons have epistemic authority with respect to what *is* contained in their self-concept and what *not*. This has to be sharply distinguished from the even more radical claim that knowledge about the truth of the self-related beliefs that actually *are* contained in a person's self-concept is *itself* a private affair – bestowing the person having the self-concept in question with epistemic authority regarding not just her self-concept, but her personality as a whole. One could summarise this in statements like, “No one can know myself better than me” or, “In the end, only I *know* what I am really like.” In order to avoid such exaggerated ideas with respect to an alleged privacy of self-concepts, one must not forget that a self-concept is a linguistic entity. The expressions persons use to characterise themselves (like “anxious”, “daring”, “musically skilled”, “striving to become a surgeon”, “being convinced that everyone can make it if he only tries hard enough”) are part of the language they share with their speech community. Consequently there are “public” (i.e. intersubjective) criteria for their correct usage (cf. our condensed account of Wittgenstein's argument against the possibility of “private” languages in Section 3.1). Of course, constituents of personal self-concepts rarely ever are so blatantly false as in the examples of an anorexic person claiming to be fat or of a person of our times claiming to be Napoleon. But for any (even the most subtle) utterance, be it self-related or not, to be meaningful, there need to be intersubjectively testable conditions under which it can be refuted. The only reason to, nevertheless, grant persons a somewhat privileged position in deciding the truth of parts of their own self-concept lies in the fact that they – in a way – have the most complete survey of the pertinent evidence for these beliefs. Since, usually, the relevant statements are unrestricted with respect to the period of time for which they claim to hold true, their empirical basis is nothing less than the sum total of what the respective person says and does. Sure enough, for each person she herself will be the only one who can claim to be *always* present when she does or says something. Therefore, it makes good sense to grant persons benefit of the doubt regarding matters of their personality as expressed in their own self-concepts. However, once there is reason to question the trustworthiness of persons, the reliability of their memory, their capacity of unbiased judgment etc., their self related statements will be scrutinised more closely. Given the many reasons for persons to lie with respect to who they really

are and the various ways in which episodic memory may delude them, the most complete familiarity with the empirical data does – just by itself – not warrant the most reliable knowledge¹³⁴.

This we have tried to make quite clear: If you want to know about a person's personality, it is not enough to just ask her about it. Still less, however, can you *do without* asking, because each and every statement persons issue about their own personality is of significance for an account of it in some way: If there is reason to assume deceptive intent, statements may indicate an image the person deliberately creates of herself. In such cases, for an adequate understanding of her personality, one has to investigate what kind of image the person sustains by the insincere statement, and in which situations and for what reasons she does so. If a statement faithfully represents a person's self-concept, it needs to be included in a comprehensive account of her personality in any case, be it right or wrong. The reason for this is not just that the ways in which a person is mistaken with respect to herself may again permit conclusions on her *actual* personality, but rather that a person's self-concept itself constitutes an important *part* of her personality, and therefore must not be omitted in an adequate account of it. Thus, any change in a person's honest beliefs on who she essentially is deserves to be regarded as a change in her personality, no matter whether the propositional contents of her new self-related beliefs are in fact true or not.

For the sake of illustration, imagine Tom who is smug about his extraordinary intelligence. At least it is part of his self-concept that he considers himself to be much smarter than most people around. Actually, the people who know Tom well would not describe him as being of above average intelligence – but who cares and who knows for sure? One day, however, Tom decides to participate in a clinical trial for a new enhancement technique that is supposed to boost general intelligence by means of a single irreversible intervention. In order to measure the effect of the enhancement participants are subjected to a number of the most carefully validated IQ tests prior and at regular intervals subsequent to the intervention. On request, the participants are told about their initial scores, but for the duration of the long-term follow-up investigation they are blinded with regard to their test results after the intervention. It is revealed to Tom that his IQ at the begin-

¹³⁴ There is some truth in saying that it is easier to assess the personality of others than one's own: If one's actions are very much out of sync with one's "ideal self", then admitting the facts about oneself can be very painful. In such situations, the subconscious temptation to rather reduce the cognitive dissonance by "rationalising" one's actions may be hard to resist. As one usually does not have to deal with *such* obstacles in assessing the personality of others, this may very well result in more objective judgement here. Everyone knows the case of people who, while totally blind with respect to their own problems and shortcomings, nevertheless have a sharp and analytical eye when it comes to assessing the personality, problems and faults of others.

ning of the study (with respect to different tests) is slightly *below* average – and it doesn’t change due to the intervention, for he is a member of the placebo control group. In the end, the only lasting result of the trial is a change of Tom’s self-concept, because after the disappointing disclosure of his initial test scores he no longer prides himself on his intelligence. In the light of what we said about the relationship between personal self-concepts and personality, Tom’s case should be described as follows: Given Tom’s IQ scores prior to the alleged enhancement, he had entertained a false belief regarding his intelligence – he thought it was above average while in fact it was not. As his scores did not change, it would have been wrong all along to describe Tom’s personality by stating that he stood out against others with respect to his intelligence. However, it would have been perfectly *right*, before the clinical trial, to refer to his wrong belief on his outstanding intelligence in describing his personality by stating that this belief constitutes a part of his *self-concept*. Since participating in the study adjusted Tom’s self-concept, his personality changed (in an advantageous way) even though his intelligence did not.

Nobody can claim to be the last authority with respect to the question who a person is compared to others – not even (as we saw) the person whose personality is actually under scrutiny. Nevertheless, given conflicting views on particular aspects of a person’s character, it is not that difficult in many cases to decide which view is true in an unambiguous sense. Questions regarding personality traits can often be settled with reference to (more or less) generally acknowledged standards like IQ tests. However, tools of that kind just offer particularly sophisticated ways to generate and analyse evidence about questions concerning personality. Ultimately, the truth of both, first *and* third person statements purporting to describe a person’s personality, must be decided in the light of how the person acts (including speech acts) and behaves. Given the complexity of this evidence, however, there is a point in saying that it will not be possible for each pair of inconsistent statements on a person’s character to single out one as definitely true (or definitely false). Nevertheless, one of these conflicting statements may still be more convincing than the other. For example, there may be reason to accept one of the statements by virtue of this resulting in a more coherent overall picture of a person’s personality.

To further flesh out this new aspect of “coherence” that shall prepare our final word on the issue of personal identity, it will be helpful to envision the way how persons express their self-concepts. Usually, a person’s account of who she considers herself to be does not assume the form of a warrant of apprehension in that she would simply enumerate traits of herself that she considers to be salient. Rather, persons convey their self-concepts in a *narrative* form by telling coherent stories of how they came to be who they are. Again, rather than merely *conveying* something they already got in their pocket, so to speak, they actually *create* their self-concepts by telling stories

about who they are.¹³⁵ Since we recognised that a person's self-concept is an integral part of her personality, *in a limited sense* persons can be said to thereby create their personalities as well. This careful wording is required as talking about a "narrative approach" may evoke associations of the seemingly limitless freedom of authors of works of fiction. To counter this impression we would like to point out again that in many ways we find ourselves *endowed* with a particular personality. This holds true for most of the p-traits we introduced above. For instance, we explicitly mentioned that talents are innate in that there is nothing a person can do to acquire one set of talents rather than another. All we are left with is to decide whether or not to make the best of our given talents. In much the same way, choleric persons do not *choose* to be overly irritable. However, they may be free to adopt different attitudes towards their tetchy nature. In general, persons can exert a reinforcing, inhibiting or shaping influence on the p-traits they find themselves endowed with by the particular way they integrate p-traits into their self-concept (especially with respect to their ideal self).

The most conspicuous difference between the narrative situation of persons developing their self-concepts on the one hand, and authors of fiction on the other is, of course, that the former are supposed to be true to the facts of their life in their storytelling. They share this particular claim of giving a faithful account of a life's course with the authors of biographies. However, as opposed to biographers, persons who give a narrative account of themselves will be expected to be true to their word in the way they lead their life after having finished telling their story – finished *for the time being*, one should add, because the stories themselves by which personal self-concepts are constituted will never come to an end for as long as a person continues to persist. In a nutshell, at each given point in time a person's former way of life,

¹³⁵ Far from new, this narrative approach to personality and "personal identity" (here in the sense of individuality) has been pioneered in different directions by a number of theoreticians, among them Taylor 1989, MacIntyre ²1985, Dennett 1992, and Glover 1988. However, as Quante (2002:22) points out correctly, personal identity over time is not the same concept as personality. Consequently, if "personal identity" is used by prominent proponents of narrativity, this is not to be understood as meaning (diachronic) numerical identity. So the way in which we address the question of personal identity over time, and thus the question of the criteria for the *persistence* of persons, by adopting a narrative approach is novel. The reason why it was not explored before presumably lies in the tacit (and premature) conclusion that the different usage of "personal identity" in prominent narrative approaches means that such a kind of approach *can not* be used to address the problem of personal identity over time. It is this tacit conclusion too which leads Quante – despite his clear understanding of the importance of biographical self-concepts (ibid.:Ch. 5) – astray: Having already predecided (ibid.:26) that problems of "personal identity" in the sense of persistence and of personality have to be strictly kept apart and solved separately, he later inevitably finds that the question of personal identity is without an adequate answer – and thus has to be substituted with a related (but different) question, namely about the persistence of human beings as organisms (ibid.:54).

as specified by her verbal and nonverbal actions as well as her behaviour, does determine the framework for credible stories on who that person is. Likewise, a person's present self-concept, and especially those parts of it declaring her projects of self-creation, determine her future actions, provided the person wants to remain credible.

Whether a person meets all the claims she (more or less consciously) raises by expressing her self-concept will be observed by the different recipients or addressees of self-constituting narratives. That these narratives do have addressees is, again, one of the noteworthy similarities to more common situations of composing narratives. However, what's special is the active part that addressees of stories through which persons express their self-concepts tend to play. Far from merely listening to what a person has to say, these addressees sometimes even qualify for co-authorship. This will mainly (though not only) be the case when persons get in a sense "lost" in their own stories about who they are. For instance, a person may be at a loss to account for some deeds of herself that do not fit at all with her self-concept. In such cases sympathetic listeners can provide support to "retrieve the storyline" by pointing out coherences the "main author" may have missed or by hinting at former actions that are in good accord with the recent disturbing ones, thereby recommending a revision of the person's self-concept in some particular aspect. In some cases, others will even take on most or all of the storytelling in place of a person whose way of being is at issue, and this is not just the case if the person in question is dead or missing: There are persons who, for a variety of reasons, will not indulge in ambitious storytelling or at least not spend many words on themselves. However, not to express your self-concept is by itself a particularly *telling* way to show character. And, what's most important: even the most reticent person, who remains imperturbably close-mouthed whenever others approach her with their unbidden characterisations, cannot avoid to confirm or refute the stories they tell about her through her deeds.¹³⁶

Because in our earlier discussion about the relationship between self-concepts and personalities we rather stressed the *critical* function the commu-

¹³⁶ It is worth mentioning this particular type of person, for in this way we demonstrate that Galen Strawson's recently raised objections "Against Narrativity" (2004) do not hit our account. Neither do we adopt the "Narrativity thesis" in its descriptive "psychological" form, claiming that all persons do look at their lives as some sort of narrative or have the habit of telling stories about themselves. Nor do we adhere to the "ethical Narrativity thesis", which would propagate self-constituting storytelling as the most auspicious way for leading a good, satisfying life. All we contend is that persons essentially have a self-concept and that this will require them *to be able* to make narrative sense of the events of their lives and to comment on stories that others may tell about them. By the way, Strawson does more than enough to prove that he himself is entertaining a self-concept in the required sense by avowing himself an "Episodic" (i.e. one of those persons "who are likely to have no particular tendency to see their life in Narrative terms" [2004:430]).

nity of persons exhibits with respect to its individual members' understanding of themselves, it was high time to finally mention also the *supportive* part other persons (friends, relatives, colleagues etc.) may play with respect to a person's efforts to come to grips with who she is. For sure, both functions, the supportive and the critical, are complementary aspects of the same process, namely the social process of interpersonal exchange in which self-concepts as well as accounts of personalities are constituted¹³⁷: By taking stock of what they did and what they experienced, persons develop a sense of individuality or even uniqueness over the course of their persistence. Most persons do occasionally or even habitually share their self-concepts with others by telling stories about what they consider to be special about themselves given their past lives and their plans for the future. In the light of what the addressees of these stories know about the storyteller's past and present way of life, they may come to their own conclusions about who that person is compared to other persons. More or less frankly and more or less sympathetically they may share their doubts or objections with the person whose self-concept is at issue. In turn, the person may either defend her self-concept so that everyone involved is eventually convinced by it¹³⁸, or she may acknowledge the observed inconsistencies and modify her self-concept accordingly. However, it is not always possible to settle dissension in a manner which is agreeable to all. If the result of interpersonal exchange on a particular person's self-concept is that it got things wrong in some particular respect, then the "official" account of that person's personality will diverge from that of her self-concept. There can be many different stories around which all purport to represent a person's personality, including the one that is told by that person herself expressing her *self-concept*. However, her *personality* is what is represented in the most *intersubjectively convincing* story of who that particular person actually is when compared to others.

5.4.5 A Narrative Approach to Personal Identity

What we have so far is a narrative approach to *personality*. So, what is it good for with respect to the problems of *personal identity*? First of all, it offers a new line of reasoning to explain the significance of episodic memory for the persistence of persons. The point is that the raw material for those stories we tell about ourselves is provided by the faculty of episodic memory. More pre-

¹³⁷ It is a distinguishing feature of narrative accounts to put emphasis on the social constitution of the self; see e.g. Taylor (1989:I.2.2.) and Quante (2002:170).

¹³⁸ We would just like to mention a few ways how a person may defend her self-concept once it is challenged in some particular aspect. Let's assume someone is pointing out an episode of my life that is at odds with how I claim to be like. First of all, I may deny the truth of his account of what I did. If I have to admit its truth, I may give some reason why I did not act in line with my general disposition in that particular situation. Or, I simply may try to outvote him by citing a number of counter-examples that support my claim.

cisely, the availability of episodic memories is a condition of the possibility of the kind of storytelling by which persons generate their self-concepts. Hence, if episodic memory is wiped out completely, as in severe cases of retrograde amnesia, then a sufficient condition for a change of personal identity to occur is fulfilled. This, of course, is perfectly in line with our earlier result that some sort of episodic memory continuity (as specified in Mem-Crit1) between person-stages is indeed a necessary condition for them to belong to the same person.

The episodic memories of persons provide the raw material for the communicative process whose narrative outcome is an intersubjectively approved account of their personalities – including a critical appraisal of their self-concepts. We consider it indispensable for an adequate understanding of persons' characters to make allowance for their narratively constituted self-concept. Only in this way one can make sure not to miss anything that is pertinent for deciding questions regarding the persistence of a particular person. By just following a highly schematised check-list in describing personalities one cannot do justice to individual variations of the contents of personal self-concepts. These variations are due to the fact that persons refer to vastly differing properties in composing their self-concepts. But even this does not account for all the variety: something quite pertinent is left out if one does not pay attention to a person's particular *narrative style* in expressing her self-concept. Some persons seem to follow the aesthetic ideal of a classic coming-of-age novel in that they see every event worth mentioning in the light of how it contributed to the eventual achievement of one overarching *telos*. Quite to the contrary, others may cultivate a thoroughly "episodic" stance towards life, avoiding the assumption of any overarching *telos*, seeing their lives as a conglomeration of independent phases which would have to be characterised by nothing so much as the contingent circumstances and tasks they were confronted with at the respective times.

It will be illuminating at this point to take a closer look at possible cases of change in a person's narrative style: Can a person who is there after a transition to an episodic outlook on herself still be considered as the same person who formerly used to think of her life as proceeding towards some *telos*? Well, why not? For instance, after some deep disappointment a person loses all her faith in a deeper meaning of life. As a consequence, she may begin to live her life from one day to the next, from year to year, thereby just changing her behaviour along with the circumstances, opportunities and roles presenting themselves to her. But it still would be *her* life. After all, she has got reasons for why her self-concept changed so radically. Others may question the consequences she has drawn from her experience of disappointment, but they will need to admit that she has got a story to tell that is worthy of consideration. As long as she cares a bit – maybe not for herself, but for others –, as long as she is ready to take responsibility for what she does and for what she did prior to her "identity crisis", there is no reason to doubt that she is

still the same person, even though profoundly changed in her personality in general and in her self-concept in particular.

We now consider a more radical change in “narrative style”, from *telos*-style narration to one which maybe could be labelled “radical deconstructivist”: When questioned about his self-concept, the radical deconstructivist will say “I don’t have a self-concept, there is nothing but ever-fleeting images of myself.” He considers “personal persistence” to be an illusion, too: Every morning he will wake up thinking “Let’s see *who there is* today”. Consequently, he will refuse taking responsibility for “his” actions. He may say, “I know someone did something yesterday, but that has nothing to do with...”, and after a pause he may continue by saying “who cares!” Well, the others will certainly care! If this “deconstructivist stance” turns out to be nothing but a superficial “attitude”, as some sort of transitory digestive trouble possibly caused by reading philosophical or spiritual literature of a certain kind, it has to count as nothing but a very weird and rather phoney sort of image in the end. The person in question will therefore be held responsible for everything he does (or did, earlier) – whether he likes it or not. If, on the other hand, his opinions are considered genuine (this presupposes that his behaviour is properly erratic), he most likely will be considered to suffer from a severe depersonalisation disorder and will be treated accordingly. This means that rather than describing the transition of the “coming-of-age person” into the “radical deconstructivist” as a change of personal identity, one would describe it as a person going so insane as to lose personhood. But why? The reason seems to be not so much that the deconstructivist story is not convincing (though in fact it is not, as it does not even *try* to account for anything, least of all the transition from the “*telos*-style” to the “deconstructivist” one). Rather, it does not count as a story in the first place (and thus the “deconstructivist style” in truth is not just one “narrative style” among others).

Quite obviously not every sequence of utterances can count as a story, but what the criteria for “story-ness” are is a harder question. There are some *minimal requirements* with respect to coherence that any narrative *with a claim to truth* (and so *a fortiori* any narrative defence or modification of a self-concept) needs to fulfil. These requirements are such that violating them in particular instances means that – with respect to the goal of intersubjective agreement – there is something not quite correct yet about the story, which has to be amended to get it “right”. Violating them more often, however, does rather result in the breakdown of “story-ness” itself.

The first requirement we want to draw attention to is that inconsistencies with respect to the time-order in which a person places the events of her story cannot be tolerated. This does of course not mean that the order in which events are to be reported has to be the order in which they actually occurred. This would be a just demand on chronicles, but not on all sorts of stories with a claim to truth. Quite to the contrary, depending on context

and purpose, such stories may present us with a whole lot of very helpful flashbacks or anticipations of things yet to come. What *can* be justly demanded, however, is that the order of events in a story with a claim to truth can be consistently deduced from it. In other words: If, according to the story, some event is both to happen and not to happen at a certain relative place in the time-line, then this is definitely something calling for amendment. It should be clear that if you have an increasing amount of inconsistencies in the time-line of a story, it will, at some point, eventually break down as a whole. Just as a bunch of “timeless” statements (like e.g. Euclid’s axioms) doesn’t make a story, a bunch of statements about events that cannot be ordered even partially is no story, either (even if terms like “post-modern”, “deconstructivist” or whatever are invoked).¹³⁹

The second requirement we would mention has to do with the observation that a mere concatenation of event-stating sentences does not make up a story either – even if the time-line is correct and the sentences are all true.¹⁴⁰ The coherence of a story thus must consist in more than just a consistent time-line: There has to be some sort of “connectedness” of its statements beyond that. To be sure, a story must feature certain recurrent objects and “themes”, but in general it would be quite futile to try getting a grip on the pertinent kind of connectedness in question by looking for purely syntactical or even semantical criteria. To get closer to the heart of the matter, it is crucial to realise in a first step that a story’s claim to truth (if it has such claim) is never just “to report something that happened”, nor “to tell it all” or “to tell it most exactly”. Rather, for every story, there is an explicit or implicit background of goals against which objects, events and levels of descriptive detail and completeness are judged pertinent or irrelevant.¹⁴¹ So, ideally, what is told in a story is relevant with respect to “what it is about”. This very general observation can be elaborated a bit further by taking into account that, even with respect to the specific goals which determine what the story is about (and what not), a story’s claim to truth (again, if it makes such claim) is never just to *report what happened*, but also in a way to *make sense of it*. This in turn can be understood in a variety of ways, the most prominent being the intentional explanation of actions and the causal explanation of other events (“Why did Stalin offer the reunification of Germany in 1953?”,

¹³⁹ Here and in the following we would ask the reader to keep in mind that we are talking about stories that claim to be true. We do not want to say that e.g. deconstructivist movies are worthless as pieces of art.

¹⁴⁰ See Danto (1985:117). His counter-example: “Naram-Sin built the Sun Temple at Sippar; then Philipp III exiled the Moriscos; then Urguiza defeated the forces of Buenos Aires at Cepada; then Arthur Danto awoke on the stroke of seven, 20 October 1961.”

¹⁴¹ Again Danto (1985:131) has a nice example to illustrate the point: “Suppose I’d wish to know what happened at a court trial. [...] I should be dismayed if [...] he were to tell me how many flies there were in the courtroom, and show me a complicated map of the precise orbits in which they flew, a vast tangle of epicycles.”

“Why did the dinosaurs become extinct?”).¹⁴² For example, a self-conceptualising story should provide us with something which a simple listing of p-traits (“social, serious, open-minded ...”) cannot provide: an answer to the question “Who are you?” in the sense entailing questions like: “How did you *become* who you are?”, “Why did you *change* in this way?” Bearing this in mind, the relevance of parts of a story can be assessed through questions like “Does this set the stage for, or contribute to the understanding of something else which the story is essentially about?”, “Is this something we can take for granted or rather something that should be explained?”, etc. Again, “ideally”, every part of a story would be relevant in the sense that it serves a function with respect to the descriptive and explanative goals associated with the story. The function of the respective part would be its *significance*, which is why we are speaking of the “requirement of significance”.¹⁴³ As with the requirement of consistency of the time-line, single violations of the requirement of significance, while sticking out like sore thumbs calling for band-aid (“Why did you tell us about that watch of your father, the one which you accidentally broke?”), will not threaten “story-ness” – but an increasing amount eventually *will* do.¹⁴⁴

The third and last requirement to be mentioned here is, actually, a generalised version of the first one: Keeping a consistent time-line means an instance of following the more general requirement of keeping truth-claiming stories free from contradictions. Therefore, the addressees of self-conceptualising stories actually have the right to object to *any* inconsistency they notice. And again we can state that, whereas truth-claiming stories are

¹⁴² To be sure, a satisfying story does not have to (and usually in fact does not) give all these answers by way of mentioning explicitly all the pertinent maxims or causal laws. However, for the story to be satisfying, they must be, in principle, distillable from it – at least to the extent of providing an explanation *schema*.

¹⁴³ To forestall any misunderstandings, it is extremely important to point out that the requirement of significance does not mean that, in their self-conceptualising enterprise, persons should try to give “meaning” to every miniscule episode or even to a major part of the episodes of their lives (this could be called a “theodicy-style” narrative). Rather, the requirement just says that significance is to be given to every episode which *is* actually making the cut, showing up in the story constituting a person’s self-concept. Note also that this does not demand or even favour *telos*-style narration (i.e. that the significance of any part of a story lies in its contribution to the same final goal).

¹⁴⁴ Though we maintain that the goal of “making sense” and the requirement of significance also hold with respect to stories that do *not* have a claim to truth (especially novels), we of course acknowledge that this may interact with other goals here. For instance, it may be extremely boring and shallow if the significance of every scene is too obvious. It even may be especially praiseworthy if the significance of many events is left unclear until the last moment, or proves to be quite different than it seemed at first (“twist”), or if the story as a whole is open to more than one “interpretation”. Again, not even a “deconstructivist story” is ruled out as a possibly subtle and worthy work of art – it is just ruled out as a *story*.

“robust” with respect to singular, isolatable contradictions, they can take only so much of these before eventually breaking down wholesale.¹⁴⁵

Of course, the internal consistency of a story is not enough to make it a true story. To this end, a story also has to be “consistent with the facts”. Regarding self-conceptualising stories, there can be more or less obvious conflict between what a person *says* and what she *does*. We already saw how to deal with “external” inconsistencies of this kind. If they can’t be “explained away” (i.e. if it can’t be shown that they just *seemed* to contradict the facts), they can be explained with reference to an image that a person may consciously create in order to fool others about her true self-concept, or with reference to a discrepancy between how a person thinks about herself and how she actually is. As a special variant of the latter case those numberless situations are worth mentioning in which persons fail to act in accordance with their “ideal self-image”. The important thing here is, though, that no amount of external inconsistencies will ever lead to a breakdown of “story-ness”. As long as it is internally consistent, a story may be as blatantly and impudently wrong as you dare to imagine – it will still be a story. So, contrary to what we found with respect to *internal* inconsistencies, the coherence of self-constituting stories, and hence the stability of self-concepts, is never threatened by *external* inconsistencies. Thus, the maniac who thinks he can take on the world all on his own *has* a self-concept, however “off-target”, while the “radical deconstructivist” has none at all.

Having dealt with structural criteria which stories (with a claim to truth) must fulfil as such (and good stories in particular), we revert attention back to our example of a person going from *telos*-style to episodic style self-conceptualisation. What are the general lessons such examples may teach us about the distinction between personality changes and changes of personal identity? Well, we came to the conclusion that persons may persist through very profound changes of their personalities *including* their self-conceptualisations. This seems only to confirm our earlier result that it is not possible to establish adequate character-based criteria for personal persistence: If all we have got with respect to two person stages at different times are two most comprehensive descriptions of personalities, including the respective self-concepts in every detail, then it will not be possible to tell whether the two person-stages they describe belong to the same person. However, by putting things this way something quite essential is missed: As we have seen by now, a person’s self-conceptualisation is not exhausted by opinions stating whether she has or has not certain p-traits. Rather, it is also constituted by narratively structured opinions¹⁴⁶

¹⁴⁵ In this we of course accept that standard logics are wrong in assuming that from a contradiction *every* proposition follows – in that case a single contradiction would suffice to make a story worthless.

¹⁴⁶ This may be a good place to state that it does not matter so much whether the “stories” a person has to tell about herself are explicit in the sense that she actually

regarding how the person acquired these traits and, consequently, by likewise structured opinions about how her self-concept *itself* has developed and about how it came to change, maybe profoundly, throughout her life. That is, a later stage of a person is connected to the earlier ones by way of narrative integration of the personalities of the earlier stages. Though this is only possible by recourse to episodic memory, it is nevertheless a much stronger link than mere memory continuity provides: In narratives, memories are arranged within a wider time-frame, and they are interlinked with respect to their content by recurrent objects and themes. By including episodic memories in a story-line or by leaving them out as “irrelevant”, storytellers value their contents, and they give a certain “meaning” (in the sense of “significance”) to them by putting them at certain, specific places within the story.

In this way a story can do several things a mere episodic memory, or even a string of such memories, cannot: Firstly, it can in a way incorporate stages which are not directly accessible by episodic memory anymore. For example, I may have no recollection at all of what I did on the 13th of February, 1984, but if I know where to place that date within the larger frame of “my story”, I will nevertheless have a pretty good idea about what kind of guy I was back then: what my temper, goals, opinions and talents were – even what I used to think of myself. The second important thing a story (as against a string of memories) can do is to tell what happened and at the same time *give an account of it, make sense of it, explain it*. This is so strong a tool that it even can deal reflectively with, and account for changes occurring over time in the story *itself*. What we have in mind here goes way beyond those trivial changes due to the mere addition of ever new parts with the advance of time: For example, a story could change in that it picks up events which were left out as irrelevant in earlier versions, or leaves out or qualifies elements which were formerly considered of special import. Maybe some seemingly inconsequential action of mine proved to be one of earth-shattering consequence much later; or what I thought was the root of all my problems was in fact important only in that it diverted my attention from my real, however repressed conflicts. In both cases, the later story would in a way explain (or try to explain) what was wrong with the earlier version, and thus would incorporate the “story of the old story” into the new story-line (“I used to tell myself that my moodiness was due to too much stress at work until I eventually realised, when my uncle died, that it had much more to do with ...”). As we have seen in the example about the person turning from *telos*-style to episodic narration, a story could even

told them to someone, or e.g. wrote them down in a diary, or rather implicit in the sense that they could be elicited by prodding her in the right way, asking the right questions etc. What counts is that she actually has a set of opinions about herself, about her life and development as a person, which together exhibit a narrative structure.

change its overall narrative style – and such a change, too, could be accounted for in the new story.¹⁴⁷

We are now in a position to state more clearly what is required for a decision as to whether two person stages (at different times) that exhibit more or less different personalities do nevertheless belong to the *same* person at different times of her life: Two person stages x and y belong as temporal parts to some (the same) person p if and only if they are continuous and the narratively structured self-concept of the later stage y does plausibly integrate stage x . Before we pause to talk about a host of issues regarding this criterion (like the meaning of “plausibly integrate”), we first propose a generalised version¹⁴⁸:

- (PI-Crit) The members of a set of person-stages M all belong as temporal parts to some (the same) person p if and only if they are all continuous to one another and for every stage whose narratively structured self-concept does not plausibly integrate all earlier stages there is a later stage which does so.

The criterion has now the form of a biconditional (“if and only if”). That is, if we manage to hold on to it, we finally got the necessary and sufficient condition for personal identity we have been after for so long. So let’s begin to discuss it.

Firstly, it may be questioned why the criterion is focussing on the self-concept of the individual. After all, didn’t we make quite a fuss about the *intersubjectivity* of the stories about personality, speaking about “co-authorship” and such, denying the individual person epistemic authority? But our focussing on the individual’s self-concept has nothing to do with epistemic authority. Rather, if the question of personal identity is put, one has to focus on certain capacities of the person (stages) under consideration, namely their capacities of memory and narratively structured self-conceptualisation. Thus, the criterion does not involve in any way that a person stage is, by definition, itself the ultimate authoritative judge about the question of what earlier stages belong to it or not, even though the objective (i.e. intersubjectively validated) answer to this question directly depends on its subjective capability of relating itself to those other stages.

¹⁴⁷ According to Danto (1981), works of art are like symbols in that they have meaning, but – unlike other symbolic representations – manage to say something about their subject matter through their mode of representation itself. In this vein, the narrative style of a story *itself* “tells” something about how the storyteller conceives of the subject matter of his story. In other words: the narrative style of a story constituting a self-concept is itself *expressing* something about how the storyteller sees himself. The content expressed in this way is therefore itself part of his self-concept (provided the storyteller could, in principle, make corresponding explicit statements – otherwise we would have to treat his narrative style as a *symptom* for a p-trait which he is not himself aware of).

¹⁴⁸ This is necessary for two reasons: Firstly, the case of two stages does not imply anything regarding the case of n stages. Secondly, only the generalised version will allow us to incorporate an important *liberalisation*.

Secondly, one may ask why the criterion does not include the liberalizing provision that later stages only have to integrate such earlier stages with regard to which they can *directly* quasi-memorise any episodes of experience from the first-person perspective. The answer to this was implicitly given above: Though episodic memory is a prerequisite for narrative self-conceptualisation, the latter transcends the former. I may not remember anything I experienced on a certain day, but as I can locate that day within the wider frame of my life, I nevertheless should have a concept of “who I was” at that time – what my temper, goals, beliefs and talents were like, what I thought of myself, and even which opportunities and hardships I was confronted with that I think were bound to have an impact on my further “story”. This is a relation I should by some measure have to *all* earlier stages of mine, even to those with regard to which I am not able to directly recall any specific experiential episodes.

Thirdly, one may wonder whether, on the other hand, the criterion is too liberal. Why don't we require every stage to integrate *all* earlier ones? Why did we add the provision that, in case of failure, things are “saved” by a later stage doing the integrating? Now, let's assume we have consecutive person stages *a*, *b*, *c* belonging to the same human body *B*, and things be such that the person at *b* suffers a “crisis” in that she is not able to integrate the earlier stage *a*, but then overcomes the crisis at *c* so that she can integrate again both *a* and *b*. Now, had we formulated our criterion more strictly, we would face the following situation: *b* and *c* would belong to one person, as would *a* and *c*, but *not a* and *b*. Therefore, we would have not one person, but *two* which just share stage *c* (and all following ones) – a case of “psychological fusion”. This would certainly be an artefact of the criterion being too strict, and it is proof of the adequacy of PI-Crit as given above that we can describe this example as the case of a *single* person recovering from a crisis.

For further illumination, let's compare the issue to the case where we have to deal with a phase of retrograde amnesia at *b* which is eventually overcome at *c*. Here, we may count the stage *b* as belonging to the same person as *a* and *c* *without* having to tamper with the definition of memory continuity, because the modal formulation “can (quasi-)memorise” provides us with the leeway we need: We already mentioned that while someone is asleep at night, he may be considered to be “able to memorise” episodes of the day before in the sense that *if* we wake him up and ask him, he will tell us something about it. In a very similar sense we may (and actually in practice do) treat a transient phase of retrograde amnesia: *If* the specific contingent obstacles that were eventually removed for *c* would have been already removed for *b*, then this stage would have remembered something. The point is that this strategy is only licit if we restrict ourselves to true conditionals which are not analytic trivialities (in the vein of “if we would have made him remember, he would have remembered”) and, furthermore, if we have an “if”-clause which is not just logically, but also empirically possible (by this provision we exclude con-

ditionals such as “if we had a *perpetuum mobile*, then we could save big on resources”). In the case of the overcome amnesia, the mere *fact that it was overcome* already provides the required backing for the assumption that it was empirically *possible* to overcome the amnesia after all.¹⁴⁹ It is important to note that this elegant way out is not transferable to the case where someone’s narratively structured self-concept is not able to integrate his earlier self, for this would amount to translate “does plausibly integrate” into “would do so if it were different”.

Now we address ourselves to the most critical question: What is “plausibly integrate” supposed to mean. First of all, the expression is to convey that it is *not* required for pieces of narrative self-conceptualisation to represent the “true story” in the sense that they would need to explain all the changes in personality in the most intersubjectively convincing way. Positively, and in the most general way, “plausibly integrate” means:

- a) there is a story (i.e. a structure which satisfies at least the basic criteria for “story-ness”) relating the earlier stages and the pertinent changes to the actually “present” stage and
- b) this story is sufficiently grounded in episodic memory to be counted as genuine.

Apart from this, the threshold for “plausible integration” will, however, not be the same *tout court*, but differ widely with context. The strongest demands will be put on the coherence of a story when there is reason to doubt that all pertinent person stages are represented by the same body. Even though *conceptually* we consider it to be possible that a person may wake up one morning in a different body, there certainly is no empirically accredited case of one person showing up in different bodies at different times. However, sometimes it so happens that a person claims to be someone who had disappeared for a long time. Especially in the past (before the advent of DNA-testing), it quite often was impossible to settle allegations of that kind by checking physical continuity between the person who had disappeared and the person claiming to be him. If the present person lays claim on the property of the person who had disappeared and if others who knew that person are affected by these claims, then we have a type of conflict on which various novels and movies are based. In situations of that kind, the claimant

¹⁴⁹ That is why cases of ongoing amnesia have to be treated with some agnostic care. For example, it is reasonable that a court does, at the one hand, *not* subject a person suffering from total retrograde amnesia to punishment for a crime committed during her “former life”. For maybe the amnesia is total, the episodic memory “wiped out” once and for all, the band of psychological continuity severed, never to be mended again. Such a person would in fact be “someone else”. However, it is also reasonable that the court will not let the amnesiac simply go free either: he will still be submitted to punishment, should the amnesia be overcome, and as long as there *is* a chance for this, the accused also has to be treated like a person with a disorder rather than a “new” person.

needs to tell a particularly convincing story about how and why he disappeared and about where he has been and what he did in the meantime. On the other hand, those who take an interest in unmasking an impostor will give close scrutiny to the requirement of sufficient grounding in episodic memory – typically by asking tricky questions the correct answers to which an impostor would be highly unlikely to know (“Your great-aunt told our investigators that you had a nickname for that sleigh you were so fond of as a child – how did you call it?”).¹⁵⁰

If, on the other hand, physical continuity between two person stages is unquestionable, there is strong *prima facie* reason to suppose that both stages belong to the same person. Therefore, the demands on the plausibility of stories by which person stages are integrated are much lower here. Actually, the onus of proof is on those who want to argue that two person stages that are represented by the same body are *not* stages of the same person. This becomes obvious when one considers court cases where a person is accused of some felony. Here we have got a profound public interest that the person who is sentenced is actually the person who committed the crime. If a defendant’s physical continuity with the criminal offender is established, then we may get one of the extremely rare instances where someone takes an interest in arguing that he emanated from a change of personal identity, so that “he is no longer the person” who committed the felony.¹⁵¹ Sometimes defendants try to argue in this vein if the crime has been committed long ago and if – in the meantime before their apprehension – their personality has changed very favourably, so that they wouldn’t be likely to commit such a felony now.

¹⁵⁰ Just for the record: Such scenarios of investigation would in principle also apply if it was (or became) empirically possible for persons to survive their bodies, be it by continuous transmogrification into different bodies or by more discontinuous means of replacement (including “reincarnation”). In the end, it all comes down to the question under what circumstances we would *accept* someone *as* someone. These circumstances are, we argue, related to someone’s capability of memory-grounded storytelling. Assuming this *is* true, then if certain phenomena became rather common (instead of just being explored in novels and movies), society would eventually begin to *treat* them as cases of reincarnation or body switching. To make this more palpable, imagine someone would show up claiming he were in fact the reincarnation of your long dead brother. Of course, you would react most sceptically, to say the least. But now imagine he would not only show the characteristic personality traits of your brother, but would also be able to recount and relate to all the things you experienced together, and could do the same, consistently, for all other members of your family. Wouldn’t you and your family eventually begin to accept him? At the very least – we dare to predict – you would do so if that sort of phenomenon, rather than being unheard of, would actually be a more common thing for which society had already developed certain traditions of coping.

¹⁵¹ The inverted commas express reservation against this wording, for if “he” is to denote the person that appears in court and if actually an identity change occurred in the meantime, then *he* never was the person who committed the felony.

However, as we saw in our discussions about “character stability” and “character continuity”, character change alone – however drastic – is by no means a sufficient condition for a change of personal identity. Even a most radical and abrupt change in character may leave a person’s capability of narratively integrating that change untouched.¹⁵² Consequently, on the basis of our narrative account of personal identity, the defendant’s best chance would be that his lawyers provide convincing evidence that he suffers from complete retrograde amnesia. In this case he will not be able to accomplish narrative integration of the person stages before onset of amnesia, thus not satisfying PI-Crit with respect to those stages. And while he will¹⁵³ not really be acknowledged as a “new person” because of this, but rather as a person with a certain form of dissociative disorder (dissociative amnesia), he would (or at least should) not be submitted to punishment until he does recover (if he ever does).¹⁵⁴ Now if the defendant does *not* suffer from dissociative amnesia, but actually *does* remember the deed, he will have a comparably harder time proving that he is not the one who is to be held responsible for it. However, even this is not impossible in principle: What the defendant would probably claim in such a case is that he remembers that “someone else” with his body committed the felony, and his lawyers would try to rack up sufficient evidence that he is suffering from DID (dissociative identity disorder, previously also called multiple personality disorder). In this case, the set of person stages associated with one and the same body can in principle be regarded as divided into several mutually exclusive subsets which are characterised by the fact that every stage *x* is only able to integrate those prior stages which belong to the set *x* itself belongs to. The subsets correspond to the different (usually *very* different) “personalities” that take turns in “surfacing” for certain periods of time. It is even more clear here than in the case of dissociative amnesia that this situation will not in fact result in an official acknowledge-

¹⁵² This also has implications for cases where, in countries with capital punishment, people sometimes change their character in a favourable way while spending years on “death row”. It is *not* a good idea to argue, for someone of that kind, that he shouldn’t be punished because he “is no longer the one who committed the crime”. For when people change from “criminal” to “good citizen”, they usually got that character change firmly integrated into their self-concept, and even make a special *point* of telling the story of that change. Thus, rather than taking recourse to an alleged “identity change”, arguments for the reprieve of these people would have to draw on the inhumanity of delivering capital punishment after years of incarceration. (Here is a last comment on the inadequacy of trying to argue for identity change through character change: If drastic character change would *indeed* all by itself count as a change in personal identity, then this would have to be acknowledged also if such a change in character had taken an *unfavourable* direction!).

¹⁵³ By most courts, in all countries, as far as we know, and according to extant law.

¹⁵⁴ It is known that retrograde amnesia can be caused by traumatizing events, and the felony *itself* could be just such an event. However, in such cases there is also a good chance for the memory eventually to return.

ment of “several persons” – DID is a disorder after all. Instead it will (or should) result in suspension of punishment.¹⁵⁵

The two examples given above (potentially different bodies with a claim to personal identity vs. one and the same body with a claim to a change in personal identity) were supposed to make clear that, and which way, the criteria for “plausible integration” of person stages in identity-preserving stories can systematically vary with the general type of situation and claim we are confronted with. Both examples were set in the context of a “court trial”, though, because we wanted to show that our narrative account of personal identity is not in danger to relegate questions of personal identity into the realm of subjective opinion (like matters of taste), but, on the contrary, can be applied to and makes a difference in situations where we want to settle such questions according to objective (i.e. intersubjective) standards of truth. On the other hand, questions of personal identity obviously do not only occur and play a role in situations where public interest is at stake. For example, a marriage may come to an end because one of the partners insists (even in these words) that the other is “not the person anymore she fell in love with” – and maybe the partner even agrees to that. Or, to give another example, some listeners may find it fully acceptable when a person accentuates a profound change of his personality by changing his name from Saulus to Paulus, and explains the change with reference to the “epiphany” of another person that passed away some time earlier. These examples show that the narrative account is also at work if there is no public interest at stake – just that what is considered “plausible integration” is then defined to a large part by more “particularistic” standards agreed by storytellers and recipients in particular narrative situations.

This should suffice for now to establish the general framework of the narrative account of personality that in our view is most apt to resolve questions concerning personal identity. We don’t need to summarise it at this stage, as this will be done in the next section. Also, its consequences for the normative issues at hand will receive more attention there.

¹⁵⁵ In fact, things can be really complicated with DID. For instance, sometimes the different “persons” associated with the body of a patient with DID do not know of each other and, consequently, can’t recall events that are experienced by those “other persons” – in which case we have to deal with forms of amnesia too. Sometimes they only partially or asymmetrically recall each other. If they do, they may have opinions about and emotions towards each other, and sometimes they seem to be able to “communicate”. Furthermore, with the progress of time, some “persons” may dissolve and others “arise”. If the courts would take it on themselves to treat DID-phenomena as anything *other* than a disorder, they would have a hard time to disentangle things. And even if they could, there would result a host of practically unsolvable problems. For example, if the person who committed the felony is still among the set of alternating persons, how could one punish *this* person without at the same time unduly punishing the others?

5.5 Summary and Consequences

In this last section of the chapter we will deal with the import the results of our analysis of the concepts of person, personality and personal identity have for a normative evaluation of the techniques of intervening in the brain presented and discussed in this book. Finally unburdened by the need for argumentative backing and discussion of rival theories, we will briefly summarise the main results of the foregoing sections and point out, for each result, *why* it is important.¹⁵⁶ On this background, we will eventually arrive at some specific recommendations to be included in the last chapter of this book.

5.5.1 Summing Up

But first things first: Let us turn to the promised summary of the main results achieved so far and the exposition of their import for the normative aims of this project. Our investigation started with looking into the concept of identity in general. Identity is a purely logical relation (i.e. definable within logic itself): objects *a* and *b* are identical if and only if all properties of *a* are also properties of *b* and *vice versa*. This raises a first non-trivial conceptual problem once *continuants* are taken into account, i.e. temporal objects that can change through time. For if an object *O* undergoes a change at time *t*, then *O* at $t'' > t$ will not share all properties of *O* at $t' < t$. How can we, on the background of the definition of identity, then still coherently talk about encountering the same object *O* at *t'* and *t''*? Given that a presuppositional analysis of the conditions of the possibility of experience shows the very concept of object permanence to be indispensable, the only solution to this problem is to adopt the position that in talking about continuants any predicate is implicitly relating the continuant to a point in time, such that the *same* object can be *Q* at one time *t* and *not Q* at another time *t'* without violating the definition of identity. This in turn entails that continuants are to be conceived of as extended in time and, consequentially, spatial continuants (i.e. “material objects”) *both* in space and in time. In other words: material objects have not only spatial, but also temporal parts, so called *stages*. To avoid certain logical confusions (especially, but not only, pertaining to cases of “fission” or “fusion” of continuants), the question whether an object *O*₁ encountered at *t* is actually the same object as an object *O*₂ encountered at *t'*

¹⁵⁶ This procedure will especially benefit those readers who are – for whatever reasons – not interested in the arguments themselves, but would just like to learn about our general train of thought, the claims we come to maintain and their connection to our eventual recommendations. However, even the mere *results* of the philosophical discussion cannot be just “listed”, but have to be arranged in an order that reflects their place within the systematic structure of questions, arguments and results. Therefore we must warn *urgently* against any temptations to mistake the raw sketches and outlines of argumentative structures to be found in this section for the arguments *themselves* establishing the presented results.

has to be conceived of as the question of whether the *stages* encountered at t and t' both belong to one (the same) object as temporal parts. To answer this question, one needs criteria: the criteria of identity over time for the sort of continuants in question (where the criteria can be different for different sorts of continuants).

The concept of a continuant as an object related to points in time logically allows for continuants to survive change, but as we do not want them to have to persist “eternally” (i.e. never ceasing to persist, but just ever changing) either, the criteria of identity over time have to be such as to also allow for an answer to the question of whether a continuant O at a certain time t has ceased to persist rather than just changed. In a first step this is achieved by means of *constitutive predicates*, those predicates by which objects are identified and re-identified. For instance, “this cup” may undergo various changes, but if a change results in undermining the conditions of ascribing the (here constitutive) predicate “cup”, then “this cup” has ceased to persist rather than changed. It follows that the persistence criteria for continuants will vary with the predicates serving as constitutive, and hence the continuant “this cup” will not be identical to “this amount of porcelain” – even if “this cup” at a certain time consists of that amount of porcelain. An intuitive way to phrase the persistence condition for a continuant that is based on its constitutive predicate is to say that, for the continuant to persist, its constitutive predicate has to “keep on applying”. However, to avoid a certain sort of circularity built into this phrasing, the exact formulation of the criterion has to be more technical: With respect to any given continuant-stage that exemplifies some continuant through a constitutive predicate (e.g. “this cup”), only such later continuant-stages to which at least the constitutive predicate (i.e. “cup”) applies can be considered as possible continuations of *that same* continuant. From this formulation it immediately becomes clear that there must be stronger criteria for really *fixing* the identity of continuants over time. For many (though not for all) sorts of embodied continuants, these criteria will conceptually involve the identity of the subsisting bodies over time, and thus their spatiotemporal continuity. However, even this will in many contexts only provide a *necessary* criterion for the object in question to still persist as “the same”, and so – depending on the sort of continuants in question – there may be various reasons to formulate even stronger criteria. For example, continuously upgrading a computer by exchanging all its parts one by one (thus ensuring spatiotemporal continuity between stages) may result in us wanting to maintain, eventually, that it is *not* literally “the same” computer we originally started out with, even though, at all intermediary points in time, the constitutive predicate “computer” actually applied to the system.

The reason why these results regarding the identity of continuants in general are of utmost importance with respect to our project is the fact that *persons are continuants*: they can (and indeed are expected to) change through

time while still surviving as the persons they are. It immediately follows that the specific criteria of *personal* identity over time have to be clarified in order to be able to assess possible risks for personal survival involved in certain techniques of intervening in the brain. This is even more urgent on the background that – as the persistence criteria for continuants vary with the constitutive predicates employed to identify them – the criteria for the identity of “this person” may very well *not* be the same as those for “this human body”. In other words: the mere fact that an intervention does not touch upon the identity of a *human body* is not in itself sufficient insurance that, after the intervention, it is still the same *person* we are dealing with. What’s more, even the continuous applicability of the constitutive predicate “person” will not by itself ensure this, for the criteria for personal *identity* have to be stronger than those for personhood.

Still, before the issue of the criteria of personal identity can be resolved, the criteria for the predicate “person”, i.e. the criteria of personhood, have to be made explicit. Rather than to just enumerate ad hoc certain capacities that are commonly associated with personhood, a more systematic and integrated approach of delimiting the range of pertinent capacities is in order. The starting point for this is that, since the times of Locke, the role of the concept “person” is that of a moral rather than just a descriptive category. Basically, in the philosophical debate since Locke “person” is used to classify those living beings which can be held responsible for their actions (which is why the concept is pre-eminently encountered in contexts of rights, duties and law).

This is, of course, *very* important since it entails that the concept of a person is *not* coextensive with the concept of a human: There can be persons that are not humans and humans that are not persons. As was stressed earlier in this chapter, the latter does of course not mean that such humans – say babies – have no rights. On the contrary: as “person” is tied to “responsibility”, babies, for instance, will just have no duties. In this context it is helpful to realise that the concept “human”, too, is not a purely descriptive category (as “*homo sapiens sapiens*” is). However, the concept “human” rather conveys rights than duties. Hence we speak of “human rights”, not of “person rights”. Compare this to the concept of a “legal person”. The main reason for widening the concept “person” so as to include abstract entities, e.g. companies, is *exactly* that this implies that they can be held responsible and, for instance, be sued. Babies, on the other hand, cannot be sued, prosecuted or convicted. Apart from these issues, it is of utmost importance to realise that for a serious consideration of concerns regarding the identity of persons which undergo a new kind of medical intervention it is *indispensable* to not just regard “person” as another word for “human being”. For on such a basis, it would be true by mere linguistic convention that *nothing* that is done to a person’s psyche could negatively affect her identity, as no matter how thoroughly one changes the personality or character of a human being it will still remain the same individual of the species *homo sapiens sapiens*.

Now, for someone to be a person, i.e. to be held responsible for their actions, they must fulfil at least the cognitive prerequisites for acquiring an understanding of the concept of responsibility: the ability to *regard oneself as oneself* and so to *attribute one's actions to oneself*. That is, a person does not only have *consciousness* (certain mental states and capabilities which are presupposed by the sheer ability to act), but *self-consciousness* (having a concept of oneself as the one *having* these mental states and *doing* those actions). That a being which is a person must have a concept of *itself* in turn presupposes cognitive capacities on its part to the extent that it is a subject of experiential knowledge and a participant in a community of language users. From this follows the non-cognitive (but nevertheless important) requirement that a person must be locatable within the common objective world. In the end, all these results may be regarded as summed up quite nicely in Locke's famous definition of a person as "a thinking intelligent Being, that has reason and reflection, and can consider it self as it self, the same thinking thing in different times and places".

Given these quite general results, and shifting focus to the observation that persons, to be responsible for their actions, require *free will* in the sense that they are able to *refrain from actions* as well as able to *act out decisions to act*, further and more specific cognitive, emotive and motivational prerequisites for personhood can be established through presuppositional analysis. In particular, persons need to be endowed with *discriminative abilities* (perception and recognition), *episodic memory*, *learning abilities*, abilities of *language* and *deliberation*, with *purposes*, a disposition to satisfy their *natural needs*, and with *sensations of like and dislike*. Apart from these "mental" requirements, it can be established that persons need to have a *body*.

In a way, the reason why these results are of importance regarding the evaluation of techniques of intervening in the brain is quite trivial: It should be made sure that their application does not – by way of side effect – eradicate any of those capacities constitutive for personhood. However, a more subtle point is that the techniques should also, at least ideally, have no negative effects on these capacities *below* the threshold of utter eradication, because such effects would have to count as "undermining" personhood. So, if with respect to an intervention technique there are certain possible side effects of that kind, this deserves special consideration in risk/benefit-weighting.

Just like the debate on the concept of a person, the modern debate on the question of personal identity can be traced back to John Locke. Since his times, the question of personal identity is the question about the criteria of persistence of particular continuants, namely persons. Even though the origin is often forgotten, talk of "personal identity" within psychology and everyday life descends from philosophical debate. In the contexts of morals and jurisdiction it is more or less obvious that the question of personal identity is important, as we attribute actions to persons and so we have to make sure that the persons we praise or condemn or convict are the same persons

that carried out the actions we praise, condemn or convict them for. However, the issue of personal identity is also of special interest with regard to the project pursued in this book, because *we* – i.e. the community of persons – *ourselves* are the objects under investigation. Thus any issues pertaining to the question of whether techniques of intervening in the brain could bear a potential risk of undermining the conditions laid down by the criteria of personal identity and persistence will be relevant insofar as persons are interested in (and have a right to) their own continued survival.

The first step towards establishing such criteria is the insight, following from the discussion of the role of constitutive predicates for the identity of continuants in general, that the continuing applicability of the constitutive predicate “person” is a necessary condition for a person’s persistence. However, it is also obvious that fulfilment of this condition alone is not sufficient to establish that it is really the *same* person we’re dealing with at different times. Candidates for the required additional criteria determining personal identity are usually categorised into two groups, “physical” and “psychological” criteria. In the discussion of the previous section it turned out that the physical criteria (namely personal identity as bodily identity or brain identity) are too strong and therefore not adequate. Although, at first blush, it may seem so, it would be a mistake to think that this result has no direct bearing on the normative dimension of the debate about present techniques of intervening in the brain.¹⁵⁷ On the contrary: As nowadays most people (quite correctly) assume that their psychic states and functions are not independent of states of their body, especially their CNS, there is an intuitive fear among the general public that interventions involving neurografting or neuroprosthetics should be *per se* regarded as bound to destroy or at least gravely endanger the personal identity of those who would be subjected to such techniques (think of the “Borg” of the popular Star-Trek universe, which can be seen as a fictional manifestation of such fears). The reason why the brain is indeed germane to personal identity is that it embodies the pertinent psychic functions (cognitive abilities and personality traits). But it is of utmost importance to recognise that it is therefore rather these abilities and traits *themselves* which have to be considered with respect to the question of personal identity, *not* their contingent physical “implementation”. If the crucial psychic functions are preserved, persons would survive, as the persons they are, even massive interventions involving neurolesioning, neurografting or

¹⁵⁷ The reason why one may be inclined to think so is probably this: “Techniques” discussed in the pertinent thought experiments, like scanning and reimplementation of psychic functions, are neither available, nor does it seem likely (as of now) that they ever will. Hence, the failure of physical criteria of identity *at best* seems to show that, though empirically improbable or even impossible, a host of novels or movies employing scenarios in which personal identity is retained through, let’s say a complete transformation of body (like in Kafka’s “Die Verwandlung” [“The Metamorphosis”]), are nevertheless conceptually coherent (in contrast e.g. to stories involving time travel).

neuroprosthetics, resulting in a brain that is quite literally not the same anymore. It is hence *not* true that intervention techniques of that kind are “by their very nature” (i.e. *as* techniques that do alter structures of or connections in the brain) bound to destroy personal identity.

Of course, this still leaves open the question of which psychic functions are the ones crucial to personal identity, i.e. the question of what are the “psychological” criteria of personal identity. Unfortunately, the most prominent candidates – criteria of psychological *continuity* – also encounter insurmountable difficulties. At most it can be shown that continuity of *episodic memory* is indeed another *necessary* criterion for personal identity. While this is undoubtedly a major result insofar as it clearly implies that intervention techniques have to be checked against possible amnesic side effects, it can be shown that continuity of episodic memory does, on the other hand, not provide a *sufficient* criterion for personal identity. Even worse, further analysis shows that the “natural” idea to upgrade the criterion to a necessary *and* sufficient one by including further reference to continuity or stability of *personality traits* (of the categories temper, motifs, propositional attitudes and talents) leads nowhere. On the contrary, it turns out that continuity criteria must always fail as sufficient criteria for personal identity (while “stability criteria” always fail as necessary criteria).

Therefore another approach to personal identity has to be found. To this, the key lies in the insight that the personality of a person is not merely the aggregate of all her p-traits as they may be described from a third person’s perspective. This aggregate view of character is valid only for non-personal beings that are incapable of holding and communicating opinions of their own with regard to the way others describe what’s special about them by ascribing character traits to them. Persons, on the other hand, are self-referential beings, who not only *have* character traits, but, moreover, hold a set of beliefs *about* these traits as well. What persons believe about their own p-traits is part of a more complex system of self-related beliefs and motifs that constitutes their *self-concept*. As self-referential beings, persons try to make sense, on their own behalf, of their actions and behaviour in a way that sheds light on who they are as against other persons. At any given time of their persistence persons may look back from their first person perspective on how they acted and reacted in various situations in order to identify those features in which they are unlike others. In a process of constant exchange with others like them persons weigh these bits and pieces of self-observation and self-interpretation and include the more significant ones in their self-concepts. Since any change of a person’s self-concept clearly manifests a change of her personality, the self-concept actually constitutes a part of the personality. Therefore, rather than merely *having* a personality that at most would need to be *discovered* by a person during the course of her life, persons – at least to some extent – actively *create* their personalities by composing their self-concepts.

The process through which personal self-concepts take shape bears significant similarities to the process of composing a narrative. To begin with, this comparison is helpful in expounding the structural properties that a set of self-related beliefs needs to possess for it to be recognised as constituting a self-concept. Since self-concepts are expressed in narratives that claim to be *true* (to the facts of a person's life and, especially, personality traits) they need to fulfil certain minimal requirements governing all stories that are associated with a claim to truth. These requirements are such that violating them in particular instances just means that – with respect to the goal of truth – there is something not quite correct yet about the story, which has to be amended to get it “right”. Violating them more often, however, does eventually result in the breakdown of “story-ness” itself. The first requirement is that, ideally, truth-claiming narratives should be free from contradictions in general; and, in particular, it should be possible to arrange all the events mentioned in a story of this kind in a consistent temporal sequence. Another structural requirement can be derived from the general fact that stories are told with certain goals in mind. For every story, there is an explicit or implicit background of goals against which objects, events and levels of descriptive detail and completeness are judged pertinent or irrelevant. So what is told in a story should be relevant with regard to “what it is about”. This can be elaborated further by taking into account that, even with respect to the specific goals that determine what the story is about, a story's claim to truth is never just to *report what happened*, but also, in a way, to *make sense of it*. To varying degrees of explicitness and detail (ranging from “elaborate” to “sketchy” and “implicit”), this will include the *intentional* explanation of actions as well as the *causal* explanation of other pertinent events. Ideally, every part of a story would be *significant* in the sense that it serves a function as regards the descriptive and explanative goals that are associated with the story. Consequently, the relevance of parts of a story can be assessed through questions like: “Does this part of the story set the stage for or contribute to the understanding of something else which the story essentially is about?”, “Is this something we can take for granted or rather something that should be explained?”, etc. Self-constituting narratives are told with a particular goal in mind, namely to convey how the storytellers became the persons they conceive themselves to be now. If someone is unable to distinguish the significance of their self-related beliefs with respect to this goal, they obviously lack a clear-cut concept of who they are.

Whether someone's self-related statements satisfy these requirements so that they constitute a coherent self-concept can be decided without knowing the actual truth of any of these statements. Of course, the internal coherence and consistency of a story is not *enough* to make it true. To this end, a story also has to be “consistent with the facts”. Once we know more about a person than what she tells us about herself, we will therefore start scrutinising the whole of her self-related statements in a different way: By drawing on imme-

diate acquaintance with her actions or on independent reports about what she did in different situations we have a chance to question the truth of parts of a person's self-constituting narratives. A person's way of describing her past development may be distorted by inaccuracies of memory or by deliberate efforts to gloss over the past. In much the same way, inconsistencies between a person's account of who she is and her actual behaviour can be explained either by her misapprehension of what marks her off from others or with reference to some image that the person intentionally created to conform to others' expectations. Finally, a person's statements about her future projects of self-creation can be proven wrong by failure to put them into action. The final result of the social process in which inconsistencies between the stories a person tells about herself and her actual way of being are first identified and then explained along these various lines is an *intersubjectively recognised account* of who that person is: A person's personality is represented by the most coherent story that persons could agree upon with respect to who she is, given the available knowledge about her past and present life, including her plans for future personal development.

According to this narrative account of personality, not only changes in a person's actions and behaviour affecting *temper*, basic long-term *motifs*, basic long-term propositional *attitudes* and *talents* need to be regarded as changes of her personality, but also changes in her narratively structured self-concept. However, none of these different types of personality changes necessarily amounts to a change of personal identity. Rather, a person is persisting as the person she is if and only if she is (not always immediately, but eventually) able to plausibly integrate personality changes (including changes of her self-concept itself) into her ever developing self-concept. "Plausibly integrate" does *not* mean that the narratively structured self-concept of a person has to be a true representation of her personality, but (to put it in the most general way) that there *is* a "story" relating earlier person stages and the pertinent changes to the actually "present" person stage and that this story is sufficiently grounded in episodic memory to be counted as at least genuine.

Apart from this, the threshold for "plausible integration" will, however, differ widely with context. To address the most problematic case first: the strongest demands would be put on the coherence of a story when there is serious reason to doubt that the person stages in question are represented by the same body. In our present day societies, reasoning about personal identity is strongly guided by the intuition that physical continuity is necessary for personal persistence, and even though in Section 5.4.1 this could not be established as a *conceptual* truth, it may well be true *empirically*. Anyway, if two person stages are *known* to not belong to the same body, then only the most coherent and detailed story – one that would pass closest scrutiny regarding the requirement of sufficient grounding in episodic memory (typically tested through tricky questions that an impostor would be highly

unlikely to answer correctly) – would have a chance to eventually lead us to accept the identity claim despite of this knowledge. (Probably, that claim would not even be considered as a residual possibility, if the “case” at hand was to remain a singularity.) If, on the other hand, physical continuity between two person stages is unquestionable, there is strong *prima facie* reason to suppose that they indeed belong to the same person. Therefore, in such cases the demands on the plausibility of stories by which the earlier stage is integrated by the later one are so much lower that, actually, the onus of proof is on those who would argue that the two person stages do *not* belong to the same person. According to our narrative account of personal identity, only complete retrograde amnesia and DID would make strong cases here. However, because there is usually at least a *chance* of recovery, these are (reasonably) treated, in practice, as cases of *one* person having a *disorder*, and *not* acknowledged as two or more persons in succession “occupying” one body. (Things would probably be handled differently if such cases were more frequent, stable and irreversible.)

There are several points with respect to which the narrative account of personal identity is of import for the evaluation of techniques of intervening in the brain: First of all, the narrative approach offers no grounds to state *in advance* what changes of his or her personality a person may successfully integrate into a coherent identity-preserving narrative. Not even on the basis of the most comprehensive knowledge of a person’s personality at some given point in time is it possible to come to any *prognoses* that a certain change of personality *must* result in a change of personal identity. Secondly, the narrative account implies that any changes in personal identity an intervention may cause would *never* go unnoticed, but would become openly manifest as severe dissociative disorders. This consequence deserves particular emphasis because it should assuage public fears that intervention techniques possibly could cause identity changes that may not be detectable. Such fears are usually based on dualistic intuitions paired with the view that scientists can only access “the Objective” (behaviour and other publicly observable data), but not “the Subjective” (the mind, the person itself).¹⁵⁸

¹⁵⁸ Fears regarding the possibility of unnoticeable identity changes even receive (probably inadvertent) theoretical backing by an influential philosophical position: the “simple view of personal identity”, according to which persons are immaterial thinking substances. As a consequence of their dualism, proponents of the simple view claim that empirical data may at best provide “evidence” for matters of personal identity, but *never* criteria in the strict sense. The simple view therefore allows for the logico-semantic possibility that a person (the immaterial substance associated with a human body) may be replaced by someone else (another immaterial thinking substance), and even for the possibility that the person vanishes, leaving the body behind as an organism still living, but no longer associated with a person (immaterial substance). For a refutation of the simple view, see Section 5.4.2.

5.5.2 Consequences

On the basis of our exposition of the normative import the results of this chapter bear on the discussion of intervening in the brain, we are finally able to formulate and justify certain ethico-political recommendations regarding research on, and application of the techniques involved. For this set of recommendations we will not claim completeness – not even with regard to issues pertaining to personality or personal identity. The recommendations will be *specific*, though, in the sense that they pre-eminently deal with such issues, and they will be *general*, at the same time, in the sense that they do not deal with any intervention technique specifically.

By the very nature of recommendations directed to a field of research on, or application of certain techniques, these recommendations will for the most part involve the specification of situations where a certain *caution* is in order. Let us make it clear from the outset: Our general stance regarding the techniques presented and discussed in this book is positive, and this should not come as much of a surprise. After all, if there were not such promise in this field, without the prospect of finding new and more effective kinds of *treatment* for old diseases and disabilities, this whole study would most likely not have been undertaken.

We are in favour of the continuation and even extension of the funding of the *research* done in the field under consideration. In our opinion, the therapeutic benefits alone suffice to warrant this statement, regardless of how one may think about the issue of enhancement. The argument for this is rather trivial and well-known: As *everything* can be misused, the mere fact that something *could* be misused towards unjustified goals cannot count as an argument against it (or we would have to object to even the most banal artefacts, such as hammers). With the question of funding in mind, one may think there are certain fields of research on techniques of intervening in the brain which are more treatment-oriented, while others are more enhancement-oriented. But, in fact, this is not the case. Rather, it seems that the same techniques *mutatis mutandis* can be used in situations of treatment as well as enhancement.

The aforementioned general positive stance on the new techniques does not mean, though, that our approval is without reflection, or uncritical. This will become evident from the recommendations to follow, which deal with the detection and prevention of certain *side effects*.

A side effect is any unintended effect of an intervention (that was carried out correctly). The purpose of the proviso “carried out correctly” is to terminologically distinguish side effects related to a technique as such from the effects of *accidents* that may occur in an attempt of its application. If, for instance, a surgeon performing a prostate operation accidentally cuts the patient’s ureter, this would not be classified correctly as a “possible side effect” of the respective operating technique. While – for any given surgical

technique – certain accidents may occur contingently (e.g. when the scalpel of the surgeon slips because of a sudden cramp in his hand), there may also be a considerable *probability* for a certain *type* of accident to occur when applying the technique in question. With regard to risk/benefit weighing, such accidents are, for all intents and purposes, treated just like possible side effects in the strict sense of the term. Consequently, for the sake of clarity, we are going to address these types of possible accidents under the general heading of “side effects”, too (to avoid cumbersome differentiation as “possible side effects *and* possible accidents significantly correlated with the application of a certain intervention technique”).

Because a side effect is an unintended effect, the decision about what is or is not counted a side effect will depend on the goals *and* there may be positive (welcome) side effects of an intervention as well as negative (adverse) ones. In the following, we will not discuss positive side effects any further, and we will also just consider such side effects which are considered negative with respect to the whole range of morally and legally licit goals of intervention. These side effects can be broken down into sub-classes in a variety of ways. For the following we propose:

- biological death
- somatic diseases (e.g. ulcer, cancer)
- motor or sensory disabilities (e.g. paralysis, deafness)
- psychic (i.e. cognitive, emotive and motivational) diseases and disabilities (e.g. depression, schizophrenia, aphasia, amnesia, sociopathy)
- personality changes (i.e. changes in temper, motifs, propositional attitudes, talents and self-concept)
- psychological death (i.e. loss of personhood or identity change)

Obviously, not all of these classes are mutually exclusive. For instance, certain somatic diseases can underlie psychic dysfunctions; a sensory disability is – via the cognitive aspect of perception – also a psychic disability; psychological death will be accompanied either by grave psychic disabilities or personality changes, etc. However, as rough as it may be, this classification should still be of some use.

Furthermore, our classes of side effects do not specifically address interventions in the psyche, but can also be applied (regarding the intended effects) to somatic interventions. There is a wide range of well established surgical procedures or medicinal drugs coming with a non-negligible – but, with respect to risk/benefit weighing, acceptable – lethality rate (death), infection risk (organic disease), or risk of other complications, which, like e.g. apoplexy, can be correlated with possible further effects like aphasia (psychic disability), loss of certain talents (personality change) or even irreversible coma (psychological death). That the subclasses of side effects given above also apply to somatic interventions should not be an objection to them, but rather a vindication: after all, the difference between somatic

interventions and interventions in the psyche is *not* that the latter deal with a different kind of substance than the former. On the contrary, interventions in the psyche (at least those we discuss in this book) are carried out as somatic interventions too. The only difference is that interventions in the psyche are such somatic interventions (usually in the CNS) that are conducted with the explicit *goal* to thereby influence certain psychic functions, whereas “ordinary” somatic interventions are conducted with more or less *exclusively* somatic goals in mind. In any case, the class of possible side effects is essentially the same for both kinds of intervention.

However, it is to be expected that because the somatic targets of interventions in the psyche are structures which implement psychic capacities or at least directly interact with such structures, certain kinds of side effects should be more prominent here than in the case of “mere” somatic interventions. Indeed, the risk of having a psyche-related side effect in, say, podiatric surgery is quite obviously much lower than in neurosurgery.

Finally, one may wonder whether side effects regarding personality changes actually belong here in the sense that they are always considered negative. One has to be very cautious here: We cannot, for instance, just suppose that a personality change is *per se* warranted if its direction is socially accepted or even valued. Even less does the mere fact that personality changes anyway all through life gives license to ignore or tolerate side effects on personality associated with an intervention, even if they should turn out to be on the socially acceptable side. Rather, an unintended personality change can only count as a “positive” side effect if the person undergoing the intervention would have welcomed that change *beforehand*.¹⁵⁹ But if that is indeed the case, then we were just very lucky, and that is why personality changes *qua* side effects have to appear on this list.

Before we go a step further towards recommendations, there is another, quite different, but nevertheless important way of dividing the class of possible side effects that has to be mentioned: the division of side effects into *conspicuous* and *subtle* ones.¹⁶⁰ Conspicuous side effects are those that will rarely go unnoticed even if no or little effort is made to check for their presence. Subtle side effects on the other hand are those that may go unnoticed (at least for a while) if not specifically checked upon. It is important to note that the distinction between conspicuous and subtle side effects is by no means co-extensive with that between major and minor ones, which already

¹⁵⁹ This can be tricky to test after the fact. The p-trait change itself can have an effect on a person's attitude towards that particular p-trait. Let's say a person was rather introverted before an intervention and is much more outgoing and social afterwards. It is quite possible (or even trivial) that the newly extroverted person approves of this trait, but what would be the value of such approval?

¹⁶⁰ While this distinction as such is not new, there seems to be no terminological convention governing it – calling the classes “conspicuous” and “subtle” is our own proposal.

expresses a normative evaluation of the side effects. For any form of intervention there may be conspicuous, but minor side effects (e.g. slight temporary drowsiness), and subtle, but major ones (e.g. a significant increase of the risk to develop cancer within the next twenty years, or an increased likelihood of a neurodegenerative disorder).

With regard to recommendations concerning techniques of intervening in the brain, we are interested in *subtle psyche-related side effects*, for obvious reasons: “Mere” somatic side effects do not fall within the scope of our chapter about persons, personality and personal identity. And for conspicuous side effects, whether major or minor, we usually do not need specific recommendations. If they are minor, they will be acceptable in the context of a risk-benefit calculation. If they are major, recommendations about them tend to be trivial and are heeded anyway (“If a method *M* to treat a condition *X* bears a significant mortality risk, then *M* should be avoided except in cases where the mortality risk is even higher if condition *X* is not treated by *M*!”).

Of the psyche-related side effects, psychological death is therefore, in *one* way, not very interesting. For, with the set of conditions for personhood established in Section 5.3, it is clear that, should personhood itself be wiped out through an intervention, this would surely not go unnoticed. A change of personal identity, however, *could* possibly be more subtle. However, considering what we established in Section 5.4.5, this would have to become manifest as a dissociative disorder (retrograde amnesia or dissociative identity disorder) and, as such, most probably still not go unnoticed for very long either. This leaves as subtle side effects unintended *effects on p-traits* and negative *effects on personhood-related psychic functions below the threshold of wiping out personhood*. (Because of their constitutive relation to personhood, even minor negative effects on such functions should receive special consideration.)

So, in intervening in the brain, there is a risk of achieving the treatment or enhancement goal while a subtle side effect would engender certain changes in personhood-related psychic functions or personality. On both counts, these could range from very minor to so drastic as to eradicate personhood or change personal identity (in which case they would of course not be “subtle” anymore).

On this background, the recommendations to be spelled out can be divided into two groups: a) research-oriented and b) application-oriented. For each of these groups we have one general recommendation:

Ad a) Research into techniques of intervention should place more weight on systematic monitoring of possible subtle (short- and long-term) side effects regarding psychic functions constitutive for personhood as well as personality changes. Special instruments (tests, questionnaires) may have to be developed for this purpose, or existing ones adapted.

Ad b) If there are known subtle side effects, there must be a risk-benefit weighing for each individual case. The patient’s or client’s *informed consent*

must include consent to such subtle risks. This would hold for treatment-oriented application *as well* as for enhancement-oriented application of the procedure. (The question whether enhancement-oriented applications are or are not illicit as such cannot be decided on the basis of the results of this chapter and will be discussed in the next chapter.)

To be sure, in order for these recommendations to follow from our results, we have to augment them with some normative assumptions, or we would be guilty of a naturalistic fallacy in our reasoning (i.e. having inferred normative conclusions from descriptive premises). Fortunately, we can fall back on some rather parsimonious additional normative assumptions here, for we only need to add to our results the claim that personhood and personal survival (and thus their protection) are of high positive value. We conjecture that all major systems of ethics will come to that conclusion (and so that our recommendations are in accord with a variety of moral stances which different readers may take).

The two recommendations given above would be quite sufficient if all people subjected to the intervention techniques discussed here would in fact be persons in the technical sense. But, of course, this is not the case. So before we can close this chapter, we still have to see what specific recommendations should be followed with respect to interventions in the brain of humans which are not (2) or are not yet (1) persons in the sense of being responsible for what they do (infants or adult people with severe mental disabilities necessitating guardianship). Before we address this question, we want to point out that our criteria of personhood by no means imply or even make it probable that there is an age at which everyone develops or has developed into a person. *De facto* there will be an average age, however with a rather large variance, and both the average and the variance will depend on biological and sociological factors. However, with respect to *recommendations*, we should take the two groups in the *legal* sense, anyway. The reason for this is simply that everyone else is *not* under legal guardianship and thus (until a court decides otherwise) has to be recognised as a person in the sense that they will not only participate in the decision, but eventually decide *themselves* (through giving informed consent) on whether they will undergo an intervention in their brain. Under the assumption that the normative legal distinctions are not just intended to be as adequate as possible, but also do achieve their objective to an at least satisfactory degree, we can still provide an answer to the question regarding cases 1. and 2. on the basis of the results of the philosophical discussion. The reason that this can be done, even though this chapter dealt with the concepts of “person”, “personality” and “personal identity”, is that in the following we will conceptually link the two groups to the community of persons by taking perspective on them as “not-yet-persons” (infants) and “not-yet-again-persons” (adult wards). Again we have two recommendations (one for each category):

- i) *Infants*: The key is that infants have the potential to develop into persons, and they normally do. They also have – still developing and differentiating – character traits. Both should be respected¹⁶¹: Ideally, interventions should neither affect negatively the *development of those psychic functions that are constitutive for personhood*, nor interfere with the *normal and healthy development of character*. Consequently, if certain such side effects are possible with an intervention technique, this has to be included as an important factor in the risk/benefit-weighting. While this seems to suffice for the treatment case, the enhancement case poses certain further challenges. We must keep in mind that in the case of legal infants we are dealing with situations where the subject undergoing the intervention is *not* also the one who decides about it by giving informed consent. As a consequence, more general possible enhancement-related problems (e.g. concerning distributive justice) notwithstanding, the *specific* problem in the case of enhancement of legal infants is that this will involve a decision about the talents of the infant on the part of the persons having the guardianship. This, however, may violate the categorical imperative in the sense that everyone is to be treated as an “end in itself” and never merely as a means to reach ones own ends. Also, setting the psychic traits of persons-to-be in advance may pre-decide or limit the later persons’ options regarding their life-plans, and thus intervene in their personal autonomy. Finally, it may be difficult or even impossible for a person to “identify” with her own p-traits as “hers” if she knows they were set in advance (just as it is sometimes difficult identifying with a prearranged marriage or a forced choice of profession) – see the next chapter for a more extended discussion of this topic.
- ii) *Adult wards*: Here, psychic functions constitutive for personhood already *are* disabled and virtually every intervention is seen under the perspective of treatment. However, with respect to disabled people, we have to take the following into account: The very ascription of a disability presupposes the assumption of a “normal” state, a state which “would be” if the disability were not here or would disappear. In comparing this state to the actual state, one would not just find deficits, but also aspects that are already “right” and which would not have to be tackled by the treatment. Without such a conception of a “normal” or “healthy” state one would not have any idea about the goal of possible “treatments”. In fact, one

¹⁶¹ In this case too we do not need to establish a system of ethics to warrant this normative claim. We only need the assumption (on which, again, all major ethical systems will agree) that the high positive value of personhood and personal survival is to be understood as including the value of everything that is a condition of a person’s healthy “coming to be”. Indeed, if we did not respect the integrity of the developmental potential of “persons-to-be” just as we respect the integrity of persons, we would thereby eventually undermine the conditions of the perpetuation of the very community of persons.

would not even have a reason to apply the concept of a “disability” at all – just as one would not think of an orang-utan as a “mentally challenged” person. Now, certainly, in a disabled person, a host of aspects that are *not* part of a disability will pertain to character traits. As these would be transformed into personality traits in a successful treatment, their integrity has to be respected from the outset and thus included as an important factor in the risk/benefit weighing.